# Early detection of cancer among patients presenting to healthcare services with fatigue

Becky White

*Thesis for the degree of Doctor of Philosophy in Cancer Epidemiology*

*UNIVERSITY COLLEGE LONDON*

# Contents

## Contents

# Declaration

I, Becky White, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

## Acknowledgements

I would like to acknowledge and heartily thank:

Yoryos, my principle supervisor – thank you for your excellent guidance from years of experience, for always making time to help, your focus on your team's long term career development, and support of your students' mental health needs. And my second supervisor, Cristina, for your warm welcome to the team, your insightful and supportive comments, and positivity.

Matt – who supervised two chapters, thank you for your outstanding statistical expertise, and I'm so grateful for your kind support as an unofficial mentor, your help in difficult times, and appreciation of terrible jokes.

Thank you to the other members of my supervisory team - Arturo, your data science experience and infectious enthusiasm, and Meena, for your valuable clinical expertise and great documentary recommendations.

To my current and former ECHO colleagues - Emma, Nadine, Bethany, Monica, and Ben, you are a joy to work with and seeing you in the office has brightened my days.

To my good friends Helen and Jordan - you have been with me from start to finish, through the highs and lows of studying during a global pandemic, and I will never forget your patience and encouragement. And my parents and Granny, thank you for always caring and being so proud of me!

Finally, to patients – thank you for continuing to make your data available for research, without which, these discoveries and their potential to improve diagnosis in future, would not be possible.

# Funding

I completed my PHD studies on a part-time basis, combined with my duties as a Data Manager/ Research Fellow for the Epidemiology of Cancer Healthcare and Outcomes Group (ECHO). As a member of the ECHO Group, headed by Professor Georgios Lyratzopoulos (GL), I would like to acknowledge funding through: The Cancer Research UK C18081/A18180 personal research award to GL; The International Alliance for Cancer Early Detection (a partnership between Cancer Research UK, Canary Center at Stanford University, the University of Cambridge, OHSU Knight Cancer Institute, University College London and the University of Manchester) project C18081/A31373; and the 'CanTest' Cancer Research UK Population Research Catalyst award C8640/A23385.

# List of figures

# List of tables

## Abstract

Many cancer patients are diagnosed after presenting to a GP with 'non-specific' symptoms, such as fatigue, which are diagnostically challenging as they are relatively common and can signal many conditions, including cancer. Symptoms like fatigue are not generally supported by referral guidelines for suspected cancer. Yet, GPs must decide which fatigued patients need urgent specialist referral for suspected cancer, or instead, investigation in primary care or 'watchful wait' management.

Motivated by a literature review identifying the potential to diagnose cancer earlier, I used GP electronic health records (EHRs) to examine short-term cancer risk in cohorts of patients presenting to GPs with new-onset fatigue. I assessed the risk of cancer and specific cancer sites, alongside a wide range of other non-neoplastic diseases, according to patient age and sex, and other presenting symptoms.

Overall cancer risk exceeded current UK guideline thresholds (> 3%) for urgent investigation for suspected cancer in older patients presenting with new-onset fatigue alongside other 'non-specific' symptoms, such as weight loss and specific abdominal symptoms. Compared to other diseases, cancer was relatively likely in older men (aged 80 years) with fatigue, but not women. Recommendations to prioritise investigating suspected cancer can be supported more strongly in men with fatigue than women.

Fatigue presentation alone was not strongly predictive of any single cancer. Moreover, risk of various potentially urgent diseases and those requiring secondary care referral was heightened in some fatigue presenters (e.g. stroke or chronic kidney disease in older men). When secondary care investigation is needed and the working diagnosis is unclear, referral through 'non-specific symptom pathways' such as a Rapid Diagnostic Centres could be considered.

Future research could compare the risk of different diagnostic outcomes in the presence of multiple diagnostic features (including symptoms, blood test results, and comorbidities), to better discriminate between possible diagnoses in fatigue presenters.

## Impact statement

This thesis provides much needed evidence to support the development of referral guidelines for suspected cancer for patients presenting in primary care with fatigue; a non-specific symptom that presents considerable diagnostic challenges.

It is the first research to examine risk of cancer overall in patients presenting with fatigue, including which patients are at greatest cancer risk, which cancer sites are most likely involved when cancer is present, and which other non-neoplastic diagnoses to consider. I have produced the first comprehensive electronic health record-based disease map of new-onset fatigue, and a novel characterisation of the co-occurrence of other non-specific symptoms in patients presenting with fatigue, along with the associated cancer risk.

I have disseminated the research by publishing three Chapters as journal articles in Cancer Epidemiology, British Journal of Cancer, and British Journal of General Practice. I have engaged with wider audiences through press releases, podcasts (BJGP and Imperial University (in preparation)), and talks I have given via collaborators (e.g. Cancer Research UK). Chapter 6 will also be submitted as a journal article and has been accepted as a poster at the Society to Improve Diagnosis in Medicine (SIDM) conference in July 2023.

As my thesis has direct applications to UK (NICE) referral guidelines for suspected cancer, and diagnostic guidelines for patients presenting with fatigue, I have provided results summaries directly tailored to these needs. I have also provided risk estimates for 237 diseases for men and women at specific ages presenting with and without fatigue to inform GPs as they make practical referral decisions.

This thesis is paradigmatically relevant to other research about which diagnoses to initially investigate in patients presenting with similar diagnostically challenging symptoms (e.g. weight loss, abdominal pain). Therefore, I have supported the wider research community's expansion of the research by publicly sharing all code used to manage and analyse the data, as well as symptom and diagnosis phenotypes; hence the research is fully replicable and can be used to investigate disease risk in other patient cohorts using electronic health records.

# 1. Chapter 1: Introduction

## 1.1 Chapter summary

There is considerable potential to diagnose cancers earlier in primary care, particularly those that present with vague symptoms. Fatigue is a vague cancer symptom associated with diagnostic difficulty, as it is relatively common and can signal a range of other conditions. GPs need information about the risk of present-but-as-yet-undetected cancer in patients presenting with fatigue. Evidence is sparse about which patients are at greatest cancer risk, which cancer sites are most likely involved when cancer is present, and which other non-neoplastic diagnoses to consider, yet this is critical for supporting clinical decision-making and diagnosing cancer earlier.

This PhD, supported by a literature review identifying the overall potential to diagnose cancer earlier, will use GP electronic health records (EHRs) to conduct population-level cohort studies examining the short-term risk of different cancers in patients presenting to their GP with new-onset fatigue. It will assess variation in risk according to patient characteristics (e.g. age, sex) and other presenting symptoms, and contextualise cancer risk against the risk of a wide spectrum of other non-neoplastic diagnoses.

## 1.2 Background

### 1.2.1 Potential to diagnose cancer patients earlier

Promptly diagnosing cancer in patients who present with new symptoms is crucial for improving survival(1–4) and patient experience(4,5). However, appropriately suspecting the diagnosis of cancer in these patients remains a challenge(6,7), as many cancers present with non-specific symptoms associated with a range of possible diagnoses of different severity and prognosis. This makes prompt and accurate diagnosis difficult, leading to diagnostic delays.

In patients with cancer who have initially consulted with symptoms, it is unclear whether and how much earlier the diagnosis could have been if improvements to the diagnostic process were made (6,7). Research using information from patient records on healthcare utilisation patterns before diagnosis could quantify this potential. In some patients subsequently diagnosed with cancer, consultation rates, records of symptomatic presentations, and the use of diagnostic tests or prescriptions, start to increase from baseline long before their diagnosis, particularly in primary care (8–10). This highlights opportunities to diagnose at least some patients earlier during what has been termed a 'diagnostic window', if it were possible to better appreciate such 'signals' through improvements in diagnostic processes or technologies(11). Currently, there is no systematic appreciation of published evidence documenting the presence and length of diagnostic windows in cancer patients.

To focus my research, in Chapter 2, I conducted a review of such 'diagnostic window' studies to quantify the potential to diagnose cancers earlier through primary care, and identify types of pre-diagnostic healthcare utilisation events (e.g. consultations, diagnostic tests) can define the start of the diagnostic window, and therefore have potential as early indicators of cancer. The review revealed that changes in primary care consultation rates were detectable six months before diagnosis for some patients diagnosed with cancers that are often characterised by non-specific ('vague') symptoms (e.g. multiple myeloma, lung cancer, and sarcoma), in keeping with other evidence identifying diagnostic delays in individual patients(12–15). This discovery, combined with known gaps in available evidence informing UK diagnostic guidelines for patients presenting in primary care with vague symptoms, indicated to me that a thesis seeking to quantify cancer risk in patients presenting with vague symptoms could help GPs assess cancer risk for individual patients and contribute to early detection efforts.

***See Chapter 2: To what extent does changing healthcare use signal opportunities for earlier detection of cancer? A review of studies using information from*** electronic patient records

### 1.2.2 Potential to diagnose cancer patients with vague symptoms earlier

When patients present in primary care with an 'alarm' or 'red flag' symptom for cancer (e.g. breast lump, blood in pee etc.), diagnostic management is typically straightforward. For example, in England, patients with 'alarm' symptoms for cancer should be referred to appropriate hospital specialties for urgent ('two-week-wait') investigation for suspected cancer (as per diagnostic guideline recommendations by the National Institute for Health and Care Excellence (NICE))(16).

However, only half of patients who go to their doctor before their cancer diagnosis report an 'alarm' symptom, though many will report 'vague' symptoms(17). Although there is no universal definition, vague symptoms are characterised by a low positive predictive value (PPV) for cancer, and if cancer is suspected, they do not usually give a strong indication of the likely primary cancer site. Examples include appetite loss, weight loss and - critical to the present inquiry- fatigue(13). Throughout this thesis I will refer to these symptoms as 'vague' symptoms, but they can also be referred to as 'non-specific' symptoms.

Patients who are diagnosed with cancer after only presenting with vague symptoms tend to experience diagnostic delays(13,17,18). This is likely because diagnostic management is less clear for patients who present with vague symptoms, compared to alarm symptoms(18–20). Vague symptoms are not generally supported by urgent referral recommendations for suspected cancer under UK NICE Guidelines, except for some specific patient groups and when certain cancer sites are suspected(17). GPs must discern which of these patients should nevertheless be investigated for cancer because of elevated risk associated with their demographic group or other vague signs and symptoms, and whether to refer on to an urgent ('two-week-wait') pathway for suspected cancer or to a multidisciplinary diagnostic centre ('rapid diagnostic centres' in England), or use a routine ('elective') referral route, or manage the patient in primary care using 'watchful waiting' or 'safety-netting' approaches.

Therefore, better quantification of cancer risk in patients presenting to primary care with vague symptoms could help GPs to identify and diagnose patients for whom the symptoms indicate an underlying cancer.

### 1.2.3 Fatigue

Fatigue is a vague symptom that is relatively common in primary care, being the primary complaint in an estimated 5-7% of general practice consultations(21–24). It is even more common in the general population, with 15-50% of people reporting experiencing fatigue recently (with 'recent' ranging from the last two weeks to three months)(25–27). Fatigue is a known prodromal symptom for many cancers, including lung, colorectal, pancreatic, leukaemia, lymphoma, prostate, renal, and ovarian cancers(28–36), with proportions of patients ranging from 4% to 45%, depending on the cancer site and study(31). There are many proposed mechanisms via which cancer might cause fatigue, including anaemia related to the underlying cancer, or tumour related cytokine release(37).

This vague symptom presents considerable diagnostic difficulty for various reasons. Firstly, its predictive value for individual cancer sites (e.g. for colorectal, lung, urological cancer or leukaemia) is low(29,38). Secondly, when cancer is suspected as the underlying cause, as it is a non-localising symptom, it is difficult for doctors to confidently suspect its primary organ site and appropriately prioritise or target investigations or referral pathways. Thirdly, reflecting its low predictive value for cancer, fatigue could also signal a range of other conditions(22,39–42).

### 1.2.4 Cancer risk assessment in patients with fatigue

As studies into fatigue have thus far considered specific cancer sites in isolation, the 'overall' risk of cancer (across sites) and also the spectrum of cancer sites associated with fatigue is not well understood, nor their relative ordering in terms of risk. As shown in Table 1.1, for patients presenting with fatigue (and no other 'alarm' symptom), UK National Institute for Health and Care Excellence (NICE) Guidelines published in 2015 recommend urgent referral for suspected cancer for certain patient groups and four cancer sites, where available evidence at the time of the development of the guidelines (i.e. pre-2015) showed the likely PPV of cancer exceeds 3% (28,29,43). The exclusion of other patient groups likely reflects lack of evidence as opposed to genuine lack of association, as most of the available studies used case-control designs that identified symptoms that were more frequently recorded before diagnosis in patients with a specific cancer, compared to healthy matched controls(29). The major cohort study available at the time did not include fatigue in its final analysis(44,45).

***Table 1.1. UK Guidelines for recognition and referral for suspected cancer, for patients with fatigue***

| **Symptom and specific features** | **Possible cancer** | **Recommendation** |
|---|---|---|
| Fatigue (unexplained), 40 and over, ever smoked | Lung or mesothelioma | Offer an urgent chest X-ray (to be performed within 2 weeks) |
| Fatigue (unexplained), 40 and over, exposed to asbestos | Mesothelioma | Offer an urgent chest X-ray (to be performed within 2 weeks) |
| Fatigue with cough or shortness of breath or chest pain or weight loss or appetite loss (unexplained), 40 and over | Lung or mesothelioma | Offer an urgent chest X-ray (to be performed within 2 weeks) |
| Fatigue (persistent) in adults | Leukaemia | Consider a very urgent full blood count (within 48 hours) |
| Fatigue (unexplained) in women aged 18 and over | Ovarian | Carry out tests in primary care<br>Measure serum CA125 in primary care |

*Published online by NICE, organised by symptom and findings of primary care investigations.*(43)

Because of such limitations of existing 'alarm-symptom' referral strategies, new models of care to help achieve speedy diagnostic resolution in patients with non-specific symptoms are being introduced in England and Denmark (termed Rapid Diagnostic Centres (RDCs) in England)(19,46,47) (Box 1). In addition to the existing NICE Guidelines, patients presenting with fatigue, who are nevertheless deemed by their GP to be at risk of cancer, can be referred to an RDC. There is a need for evidence to inform the development of these symptom-focussed referral pathways for vague symptoms, rather than cancer site-focussed investigation pathways for alarm symptoms.

***Box 1. Symptoms meeting criteria for referral to a Rapid Diagnostic Centre (RDC) non-specific symptom pathway according to NHS England guidance (2019)***(46)***:***

- New unexplained and unintentional **weight loss** (either documented >5% in three months or with strong clinical suspicion);
- New unexplained constitutional symptoms of four weeks or more (less if very significant concern). Symptoms include **loss of appetite, fatigue, nausea, malaise, bloating**;
- New unexplained vague **abdominal pain** of four weeks or more (less if very significant concern);
- New unexplained, unexpected or progressive **pain**, including **bone pain**, of four weeks or more;

- Optional: new and unexplained **breathlessness** for more than three weeks (not requiring admission and not due to heart failure, VTE, IHD, COPD or Chest infection);
- Optional: Unexplained **thromboembolism** (depending on local alternative pathways)

Therefore, more research is needed to enable GPs and RDCs to disentangle the possible underlying causes of fatigue, and make decisions about

- which patients presenting with fatigue to primary care need immediate referral for suspected cancer,
- and which need further investigation in primary care or management through 'watchful waiting' or safety-netting approaches (with possible referral later on if symptoms persist or worsen).

Therefore, in Chapter 4, I aimed to establish the risk of cancer conferred by fatigue presentation; the length of the period after presentation during which patients remain at increased risk; and how the risk varied by patient demographic characteristics (e.g. sex, age). Such information could help GPs to both suspect the diagnosis of cancer quickly in cases where the patient's fatigue indicates an underlying cancer, and minimise unnecessary tests for patients with fatigue who are at low risk of cancer.

***See Chapter 4: Risk of cancer following primary care presentation with fatigue: A population-based cohort study of a quarter of a million patients***

### 1.2.5 Cancer risk in patients with fatigue and other vague symptoms

While it is clear that patients presenting with fatigue who also present with an 'alarm' symptom for cancer (e.g. breast lump, rectal bleeding) should be urgently referred for suspected cancer, the referral strategy is considerably less clear for patients with fatigue either as the sole presenting symptom or in combination with potential cancer features that, like fatigue, are non-alarm or non-site specific (e.g. weight loss, abdominal pain, anaemia).

How often fatigue presents alongside other symptoms and the associated risk of underlying cancer, however, is not known. A handful of cohort studies of patients with other vague symptoms, including weight loss or abdominal symptoms have characterised related cancer risk (48–50).

Current evidence assessing cancer risk in patients with fatigue in combination with other presenting features is limited to specific cancer sites(30–32,35,38,51) or specific symptom combinations(49,52). A detailed examination of cancer risk in patients presenting with new-onset fatigue with or without other non-alarm symptoms and in the absence of alarm symptoms would support GPs to select patients for referral in a group of patients for whom diagnostic management is particularly challenging.

In Chapter 5, I therefore aimed to estimate the short-term risk of incident diagnosis of any malignant neoplasm (excluding non-melanoma skin cancer) in patients who present with new-onset fatigue without accompanying alarm symptoms for cancer, according to combinations of other presenting vague symptoms.

***See Chapter 5: Underlying cancer risk among patients with fatigue and other vague symptoms in primary care: a population-based cohort study***

### 1.2.6 Risk of specific cancer sites and other diseases

Diagnosis in patients with fatigue is particularly challenging, because the symptom has a low positive predictive value for a range of diseases(21–24). Other than cancer, diseases that are known to be associated with fatigue include, but are not limited to: coeliac disease, chronic fatigue syndrome, depression, hypothyroidism, iron deficiency, post-viral fatigue, and vitamin deficiency(22,37,39–42,53). More rarely, fatigue may indicate the presence of cancer, autoimmune disease (e.g. systemic lupus erythematosus), or chronic infections (e.g. HIV, hepatitis C), heart disease or diabetes(37,53,54). Underlying causes of fatigue can be categorised as shown in Box 2.

***Box 2. Aetiology of fatigue, adapted from BMJ Best Practice Guidance***(37)

**Cancer**

**Cardiovascular disease:** Heart failure, acute myocardial infarction, atrial fibrillation

**Drugs and toxins:** Recreational drugs, antihistamines, antihypertensives, anti-arrhythmics, antidepressants, anti-emetics, antiepileptics, corticosteroids, diuretics, and neuroleptic agents, ticagrelor, chronic alcohol misuse, heavy metal toxicity

**Endocrine disorders**: Hypothyroidism, diabetes mellitus, Addison's disease, vitamin D deficiency, hypopituitarism, acromegaly, growth hormone deficiency, hyperthyroidism, Cushing's syndrome, diabetes insipidus

**Gastrointestinal disorders:** Coeliac disease, chronic liver disease, inflammatory bowel disease, irritable bowel syndrome

**Haematological disorders:** Anaemia, chronic myeloid leukaemia, myelodysplastic syndrome, lymphoma, heavy metal toxicity

**Idiopathic causes:** Chronic fatigue syndrome, systemic exertion intolerance disease

**Infectious disease:** Epstein-Barr virus (EBV), HIV, COVID-19, Lyme disease, cytomegalovirus, toxoplasmosis, Q fever, brucellosis, tuberculosis, coxsackie B virus, chlamydia, mycoplasma, influenza virus

**Neurological disorders:** Parkinson's disease, stroke, multiple sclerosis, lateral amyotrophic sclerosis, myasthenia gravis, dystonias, myopathies

**Psychiatric and psychosocial disorders:** Depression, anxiety and somatisation disorders

**Pulmonary disease:** COPD, sarcoidosis, asthma, pulmonary HTN, pleural disease, and pneumonitis

psychosocial stressors

**Renal disorders:** Haemodialysis, renal failure

**Rheumatological disorders:** Systemic lupus erythematosus, fibromyalgia, rheumatoid arthritis

**Sleep disorders:** Insomnia, obstructive sleep apnoea/ hypopnoea syndrome, obesity hypoventilation syndrome, restless legs syndrome

Although many of these conditions are already included as potential diagnoses in UK diagnostic care guidance for fatigue(54), a quantification of these risks, and how they vary by patient age and sex, can support decisions by GPs about which potential diagnoses to consider in patients when they

initially present with fatigue. Ranking different diseases by their incidence in fatigued patients can provide information about the most likely diagnoses and consequently which tests or referral routes to consider first. A comparison against the risk of cancer would highlight whether follow up diagnostic strategies are important when cancer has been excluded. Finally, considering the risk of multiple diseases could inform Rapid Diagnostic Centre (RDC) referral guidelines regarding fatigue, because in the absence of a clear working diagnosis, patients at a generally elevated risk of multiple diseases could benefit from referral to an RDC.

No population-level study thus far has quantified the risk of multiple diagnostic outcomes in patients presenting with fatigue, although recent studies have used electronic health records to assess the risk of a small number of different diseases in cohorts of patients presenting with other non-specific symptoms (48–50) or signs of disease(55). However, these focussed on specific pre-selected diseases that were deemed serious and related to the symptom based on prior clinical as opposed to epidemiological knowledge, and no study has yet quantified the full 'disease signature' of a symptom by examining a large range of possible diagnoses.

Therefore, in Chapter 6, to support general practitioners as they assess which serious and non-serious diagnoses to consider after initial presentation with new-onset fatigue I aimed to quantify the short-term risk of a wide range of possible diagnoses.

***See Chapter 6: Risk of incident cancer compared to other diseases after presenting in primary care with fatigue: a population-based cohort study***

### 1.2.7 Methodological issues in studies of disease risk in symptomatic cohorts using electronic health records

In the development of the three empirical chapters examining disease risk associated with fatigue, I investigated and addressed a number of methodological issues involved in using electronic health records (EHRs) to conduct population-based risk studies in symptomatic cohorts. I aimed to share my learning with other researchers to support the development of future similar studies using EHRs.

Therefore, in Chapter 3, I prefaced the empirical studies with an examination of relevant methodological issues, such as defining a symptomatic cohort, identifying co-occurring symptoms, specifying a follow up period for diagnoses, and phenotyping symptoms and diseases in EHRs.

***See Chapter 3. Methodological issues in studies of undetected disease risk in symptomatic cohorts using electronic health records***

## 1.3 Thesis aim and objectives

The overall aim of this thesis was to identify which patients presenting to primary care with new-onset fatigue should be investigated for suspected cancer by GPs.

The specific objectives are summarised in Table 1.2, which also shows the Chapter where these objectives were addressed.

***Table 1.2. Thesis objectives***

| Objectives | Addressed in chapter |
|---|---|
| Quantify the potential to diagnose cancers earlier as signalled by increases in healthcare use pre-diagnosis, to inform the focus of early detection research, including this thesis. | 2 |
| Establish the risk of as-yet-undiagnosed cancer associated with fatigue presentation, and related variation by demographic group | 4, 5 |
| Establish the relative risk of different cancer types associated with fatigue presentation, to guide diagnostic strategies. | 4, 6 |
| Understand change over time in the risk of cancer diagnosis after an initial presentation with fatigue, to inform healthcare professionals and patients as to the duration of a 'safety-netting' period, within which there should be heightened vigilance of cancer. | 4 |
| Estimate cancer risk in patients with fatigue according to the presence or absence of other potential cancer symptoms. | 5 |
| Describe the disease signature of fatigue by quantifying the risk of a range of diseases, to guide diagnostic strategies. | 6 |
| Support efforts by the wider research community to harness primary care electronic health records to improve earlier detection of cancer, by discussing methodological issues when using Electronic Health Records (EHRs) in population-based risk studies in symptomatic cohorts. | 3 |

## 1.4 My contributions

I wrote this thesis, and designed and conducted the literature review and empirical studies. As a member of the Epidemiology of Cancer Healthcare and Outcomes Group, I have presented emerging methods and findings in project meetings with my primary supervisor Professor Georgios Lyratzopoulos, and my secondary supervisors Dr Cristina Renzi, Dr Matthew Barclay, Dr Meena Rafiq, and Dr Arturo Gonzalez-Izquierdo, and received feedback.

I was responsible for designing and conducting the literature review, including identifying and analysing the reviewed studies and drafting the manuscript, under the supervision of Prof Georgios Lyratzopoulos and Dr Cristina Renzi. Where needed, statistical and methodological advice was provided by Prof Gary Abel and Prof Henry Jensen, and clinical advice by Prof Georgios Lyratzopoulos, Dr Cristina Renzi, and Dr Meena Rafiq.

I was responsible for designing and conducting the empirical studies, including conducting the analysis and drafting the manuscripts. I developed the SQL code used to define and refine the cohort selection criteria, Stata, and R code used to manage and analyse data, with contributions from Dr Matthew Barclay and Ms Nadine Zakkak.

Where needed, statistical advice was provided by Dr Matthew Barclay, who also supervised the methodological issues discussion (Chapter 3) and empirical study on multiple disease outcomes (Chapter 6), and clinical advice by Prof Georgios Lyratzopoulos, Dr Cristina Renzi, and Dr Meena Rafiq.

I collected and refined symptom and disease phenotypes based on previously published phenotypes. All authors of these phenotypes are acknowledged where used in the studies, including notable contributions from all members of the supervisory team, Prof Willie Hamilton and Dr Sarah Price from Exeter University, and Dr Annie Herbert (formerly of UCL).

Specific contributions by my supervisory team and additional co-authors (Prof Willie Hamilton, Dr Sarah Price, Dr Henry Jensen, Prof Gary Abel, Dr Brian Nicholson, Prof Spiros Denaxas, and Ms Nadine Zakkak) were made to the manuscripts relating to Chapters 2, 4, 5, & 6. These were originally published as multi-author journal manuscripts (with the exception of Chapters 3 and 6, which will later be submitted to journals).

# 2. Chapter 2: To what extent does changing healthcare use signal opportunities for earlier detection of cancer? A review of studies using information from electronic patient records

## 2.1 Chapter rationale

In patients with cancer who have initially consulted with symptoms, it is unclear whether and how much earlier the diagnosis could have been if improvements to the diagnostic process were made. Research using information from patient records on healthcare utilisation patterns before diagnosis could quantify this potential. To focus my research, I aimed to quantify the potential to diagnose different cancers earlier through primary care, and reflect on the potential to incorporate measures of changing healthcare use into risk prediction studies in patients presenting with non-specific symptoms such as fatigue.

## 2.2 Publication

This chapter has been published in the peer reviewed journal, Cancer Epidemiology:

White, B., Renzi, C., Rafiq, M., Abel, G. A., Jensen, H., & Lyratzopoulos, G. (2022). Does changing healthcare use signal opportunities for earlier detection of cancer? A review of studies using information from electronic patient records. Cancer Epidemiology, 76, 102072. https://doi.org/10.1016/j.canep.2021.102072



## 2.3 Author contributions

Authors: Becky White, Dr. Cristina Renzi, Dr. Meena Rafiq, Dr. Gary Abel, Dr. Henry Jensen, Prof. Georgios Lyratzopoulos

BW, GL, and CR conceived and designed the study and agreed the search and data extraction strategy. BW identified and analysed the studies, under the supervision of GL and CR and with a sample of studies independently reviewed by CR. MR, GL and CR provided clinical input into interpretations. GA and HJ provided statistical and methodological expertise. All authors contributed to drafting and revising the article.

## 2.4 Abstract

**Background**

It has been proposed that changes in healthcare use before cancer diagnosis could signal opportunities for quicker detection, but systematic appreciation of the potential of such evidence is lacking. I reviewed studies examining pre-diagnostic changes in healthcare utilisation (e.g. rates of GP or hospital consultations, prescriptions or diagnostic tests) among patients subsequently diagnosed with cancer.

**Methods**

I identified studies through Pubmed searches complemented by expert elicitation. I extracted information on the earliest time point when diagnosis could have been possible for at least some cancers, together with variation in the length of such 'diagnostic windows' by tumour and patient characteristics.

**Results**

Across twenty-eight studies, changes in healthcare use were observable at least six months pre-diagnosis for many common cancers, and up to two years or longer for colorectal cancer, multiple myeloma and brain tumours. Early changes were also identified for brain and colon cancer sub-sites.

**Conclusion**

Changing healthcare utilisation patterns before diagnosis indicate that future improvements in diagnostic technologies or services could help to shorten diagnostic intervals for cancer. There is greatest potential for quicker diagnosis for certain cancer types and patient groups, which can inform priorities for the development of decision support tools.

## 2.5 Introduction

Promptly diagnosing cancer in patients who present with new symptoms is crucial for improving survival(3,4,56,57) and patient experience(5). However, appropriately suspecting the diagnosis of cancer in these patients remains a challenge(6,7), as many cancers present with non-specific symptoms associated with a range of possible diagnoses of different severity and prognosis. This makes prompt and accurate diagnosis difficult, leading to diagnostic delays. Information from electronic health records (EHRs) remains a rich resource for supporting the diagnostic process and targeting improvement efforts(58,59).

In cohorts of patients subsequently diagnosed with cancer, consultation rates, and the use of diagnostic tests or prescriptions are known to increase from baseline long before their diagnosis(8–10). For example, rates of primary care consultations among women subsequently diagnosed with colorectal cancer started to increase from nine months before diagnosis, compared to controls (Figure 2.1) (8). The onset of changes in healthcare utilisation rates defines the start of a 'diagnostic window', during which quicker diagnosis would in principle be possible. This highlights opportunities to diagnose at least some of the patients sooner, by better appreciating and acting on the 'signals' indicated by changing patient healthcare utilisation(8,11), or other signs and symptoms within the diagnostic window.

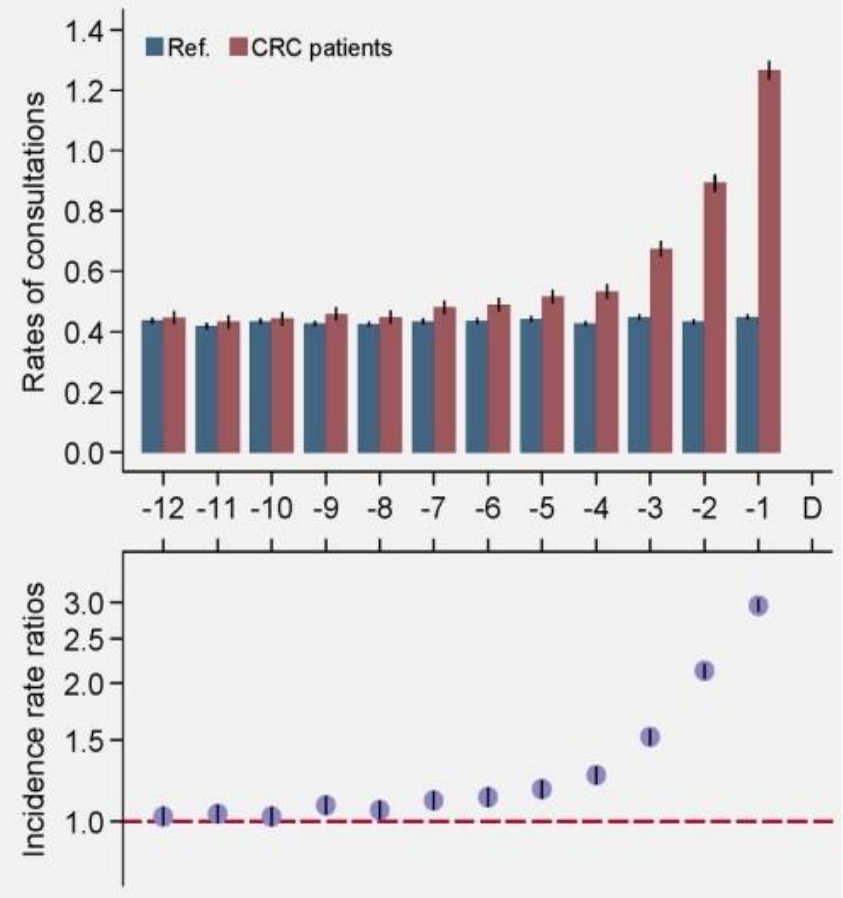



***Figure 2.1. Exemplar evidence examining healthcare utilisation changes before diagnosis of cancer.***

*Reproduced with permission from Hansen et al*(8)

Nonetheless, there is currently no systematic appreciation of how much earlier cancer patients could be diagnosed in principle, as signalled by the onset of increasing healthcare use pre-diagnosis, and for which patients' potential opportunities are greatest. Further, the exact nature of different healthcare utilisation events that could be used to identify the onset of diagnostic windows is unclear. Motivated by these realisations, I reviewed evidence from population-based observational studies reporting on the patterns and timing of healthcare utilisation events before cancer diagnosis.

I aimed to summarise the maximum length of reported diagnostic windows, quantifying the earliest point that cancers can be diagnosed as indicated by changing patterns of consultations (and presenting signs and symptoms), prescriptions, diagnostic tests (and abnormal test results) or other changes in patterns of healthcare utilisation. I aimed to identify the earliest 'inflection point' identified by each study for each cancer type, defined as the point before diagnosis when rates of a pre-diagnostic event of interest increased above a background rate (or, as applicable to diagnostic tests, when average test values changed from a background rate). I also aimed to quantify any variation in the length of the diagnostic window by cancer type, as well as describing variation by other tumour and patient characteristics.

## 2.6 Methods

### 2.6.1 Search strategy and selection criteria

Study selection followed a three-step process (Figure 2.2). In step one, all studies published before 5th July 2021 were identified for inclusion through searches of the Pubmed database. The first search was for the key terms: "*cancer[Filter] AND early detection of cancer[MeSH Terms] AND ("signs and symptoms)[MeSH Terms] OR "before diagnosis" OR pre-diagnos* OR prediagnos*)*. The second search used relevant author names identified via expert recommendation (see Appendix 10.2.2 for search terms). Additional studies were identified via expert recommendation by co-authors, and tracking citations within these recommended articles. 64% (N = 18) articles identified via expert recommendation were also detected in the Pubmed search, suggesting that the PubMed search was reasonably sensitive to articles of relevance, though it also identified a substantial number of additional papers.

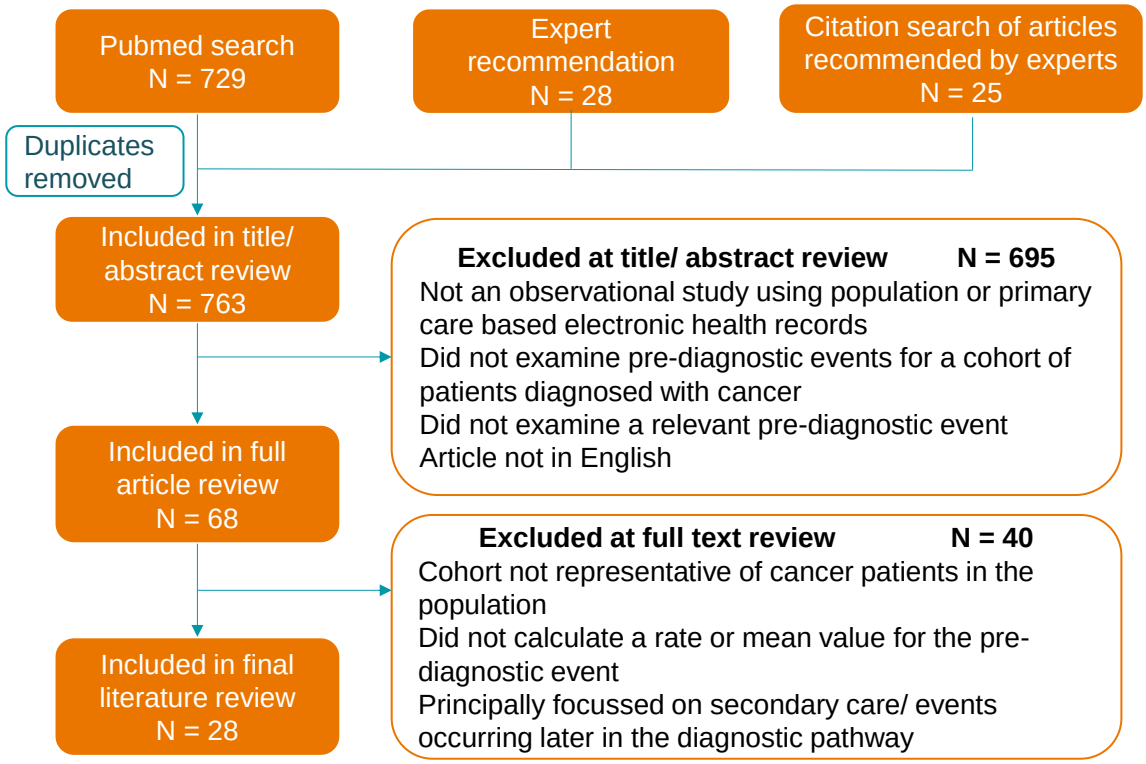



*Figure 2.2. Flow diagram of numbers of studies identified and included in review*

In step two, study titles or abstracts were included if they were observational studies using primary care or population-based electronic health records, and they studied relevant pre-diagnostic events for a cohort of patients diagnosed with cancer. Included studies should have investigated the frequency and timing before diagnosis of one of the following event types: primary care consultations, secondary care consultations, medication prescriptions, diagnostic test use (and/or related test findings), or surgical procedures in relevant specialties. These event types were determined to be broadly relevant to the early detection of cancer by the authors of this study,

based on clinical experience. There was no pre-defined list of specific relevant prescriptions, tests or surgical procedures of interest, as this would depend on the cancer type(s) examined by each study. Only articles available in English were included.

In step three, full articles were reviewed, if the cohort was representative of patients diagnosed with cancer in the population (e.g. excluding study cohorts from clinical trials, blood donor databases, or solely of patients with recurrent cancer). Studies principally focussed on secondary care patients, or events that generally occur later in the diagnostic pathway or as part of confirming a malignancy (e.g. breast biopsy, breast mammogram)(60) were excluded. To identify changes over time, studies were included if they calculated a rate or mean value of a pre-diagnostic event using suitable time intervals (i.e. studies excluding the final year before diagnosis, or treating the entire pre-diagnostic period *en bloc* were deemed unsuitable and were excluded).

Studies should have explicitly reported an 'inflection point' in the article text. If a study did not do this, but included data in figures or tables that enabled its unambiguous identification, I extracted information about the first time period (e.g. month) when confidence intervals indicated that the outcome of interest (e.g. consultation rate) was significantly different to the time period immediately before (or where applicable, to controls).

A second author repeated the selection process for a random sub-group of 70 (9%) studies identified in step one, to check for concordance in study selection. The second author made the same decision (whether to include or exclude) for 98.6% (n = 69) of the studies, and the discordant study was excluded by consensus.

### 2.6.2 Summarising evidence on the length of the diagnostic window

I summarised the range of inflection points across the studies by event type, for all cancers combined and for specific cancer sites. Where studies reported more than one inflection point for the same event type (e.g. primary care consultations for relevant symptoms only and primary care consultations for any reason) or patient groups (e.g. males and females), the earliest single inflection point for the event type was chosen. The length of the diagnostic window was defined by the number of months between the extracted inflection point and cancer diagnosis. Finally, where reported, I extracted values for the diagnostic window length by tumour characteristics (e.g. tumour sub-site, presenting symptom), patient factors (e.g. sex, age), and other factors (e.g. route to diagnosis).

## 2.7 Results

### 2.7.1 Search yield and study selection

763 studies were initially identified, of which 28 were included in the final review (Figure 2.2)(1,8–10,61–84). All but four of the selected studies were carried out in Denmark or the UK, while the four remaining studies were set in Germany, Sweden, Australia, and the Netherlands. Selected studies were published between 2010 and 2021, and included patients diagnosed with any type of cancer, and/or 25 individual cancer sites (Figure 2.3). Seven of the selected studies included children and young adults only, 17 included adults only, and four did not specify the age range.

### 2.7.2 Methodological considerations

There was variation in methodological approaches, with 11 studies using a case-only and 17 a case-control study design. Identification of inflection point timing was either based on visual inspection or statistical estimation of the point when rates among cases changed either compared to baseline (in case-only studies), or corresponding synchronous rates among controls (in case-control studies) (Table 2.1). Four studies used more than one approach for inflection point identification, yielding different estimates within the same study(75,80,82,83). For two studies, I identified the inflection point using information on estimates and confidence intervals reported in the selected studies(77,78). In one of these, study authors identified some inflection points in the commentary which were different to those I identified using confidence intervals provided graphically(77). Where the timing of the inflection point was compared between patient or tumour subgroups, no studies employed a formal statistical test.

There was variation in how the time before diagnosis was parameterised; weekly (N=2), monthly (N=18), bimonthly (N=4), and longer time units (including quarters or six-month periods, and variable period lengths) (N=7). Some studies used more than one time unit of analysis. Observation began at different points before diagnosis: 12 months (N=8), 18 months (N=3), two years (N=9), three years (N=3), four years (N=1), and five years (N=4). Overall study sample sizes ranged from 1,606(79) to 353,087(1) patients. Further, for some studies using stratified analysis, the number of patients in specific groups was particularly low, for example under 100 patients(65).

*Table 2.1. Summary of key methodological approaches used by published evidence to identify the onset of changing healthcare utilisation before cancer diagnosis ('inflection points'), and recommendations for future research*

*For more detail, please refer to the published table available at:* https://www.sciencedirect.com/science/article/pii/S1877782121001892

| Methods used by studies* to identify inflection points | Considerations | Recommendations |
|---|---|---|
| **1. Visual inspection of a time series graph to identify the time period when estimates among cases appeared to change (either compared to baseline for cases, or to controls) (13 studies)** | • Poor reproducibility compared to statistical comparisons.<br>• Qualitative description of patterns may be useful when complex healthcare utilisation patterns are present (e.g. multiple inflection points).<br>• Can identify notable/ substantial changes (as opposed to small but statistically significant changes)(64). | Consider identifying the inflection point using statistical comparisons to improve reproducibility, bearing in mind that even small changes in rates of pre-diagnostic healthcare use may result in significant findings. Correction for type 1 errors caused by multiple testing may be needed (e.g. Bonferroni). |
| **2. Statistical identification (case-only studies) of the first time period when estimates among cases were significantly different to a 'baseline' period (3 studies)** | • Sensitive to whether the inflection point is identified as the first time period that is statistically different to:<br>  ○ the start of the whole observation period (i.e. baseline), or<br>  ○ the period immediately before e.g. month by month (if changes are gradual, they may not be statistically different among adjacent periods).*<br>• Without a comparison group, changes could reflect secular trends unrelated to cancer, changes in healthcare practice, or cohort ageing effects. | Where controls cannot be selected appropriately, case-only designs could be used. Consider how the 'baseline' period is defined, plus possible underlying secular trends, changing healthcare practice, and cohort ageing effects. |
| **3. Statistical identification (case-control studies) of the first period when estimates among cases were significantly different to controls (13 studies)** | • Can be used to account for underlying secular trends and other limitations of case-only study designs.<br>• The selection of appropriate controls can be challenging(85) and imperfect matching can potentially inflate observed diagnostic window length(63). | Appropriately-designed case-control studies is can overcome limitations of case-only designs. However, simple comparisons between cases and controls in each time period could be sensitive to background differences between cases and controls Background estimates and secular trends in both cases and controls should be modelled. |
| **4. Maximum likelihood estimation of the inflection point (i.e. identifying the time period for an inflection point which provides the best fit to the data) (1 study)** | • Does not rely on statistically significant changes in estimates between individual time periods to identify an inflection point(67).<br>• Underlying secular trends, changes in healthcare practices, and cohort ageing effects can be modelled. | This approach may circumvent issues in both case-only and case-control designs. |

*For two studies, not shown here, I identified inflection points based on the estimates and confidence intervals provided. For Wang et al, I used method 2, identifying the first time period when estimates among cases were significantly different to the period immediately before*(78)*. I noted that results were different if comparing to the time period at the start of observation. For Morrell et al, I used method 3*(77)*. Four studies used more than one approach for inflection point identification, yielding different estimates within the same study*(75,80,82,83)*.*

### 2.7.3 Event types studied

Primary care consultations were the most widely studied event across cancer sites, examined by 25 studies, spanning 15 of the 16 individual cancer sites, and all cancers combined. Secondary care consultations were examined by seven studies, spanning nine cancer sites, and all cancers combined. There was heterogeneity between studies regarding the type of primary and secondary care consultations included. For example, regarding benign brain tumours, some studies included consultations for any symptom(9,83), and some only included those for specific symptoms(72) (Table 2.2), with the specific symptoms considered further varying between studies. Heterogeneity also arose from whether primary care consultations via any contact method(72), or only face to face consultations were considered(9,83). This review did not compare diagnostic window length by contact method, as studies did not present findings by the specific method (e.g. face-to-face, email, telephone).

Diagnostic test use was examined by ten studies, spanning 20 individual cancer sites, and all cancers combined. Diagnostic test use encompassed imaging tests, biopsies, lung function tests, blood tests, urine tests, pulmonary function tests, electrocardiography, streptococcal throat infection, psychometric tests, and, in one study, unspecified 'paraclinical' examinations in particular hospital specialties(68). Only two studies examined changes in diagnostic test findings, encompassing mean values for various blood tests, among patients subsequently diagnosed with multiple myeloma(10) and lung cancer(74). Prescriptions were examined by four studies, spanning three individual cancer sites, plus all cancers combined. Surgical (i.e. orthopaedic, dermatology, and plastic surgery) procedures were only examined by a study of sarcoma patients(68).

### 2.7.4 Longest reported diagnostic windows by cancer site

For studies considering the outcome of all cancers combined, the length of the diagnostic window ranged from six(1,65,71,75) to 16-18 months(64,80) before diagnosis, as inferred from a detectable change in pre-diagnostic event rates or test values. The length of the diagnostic window varied substantially by cancer site (Figure 2.3, Table 2.2).

The longest reported diagnostic windows (for all patients or at least a subgroup of patients) were four years (43-48 months) pre-diagnosis for all brain tumours combined(82), three years for multiple myeloma(10), and two years for colorectal cancer(8,69,70,79), and benign brain tumours(65,76). Diagnostic windows of between six to 12 months were reported for lung cancer(63,74,77), sarcoma(68,83), bladder and kidney cancers combined(67), childhood/ adolescent lymphoma and leukaemia(65,83), malignant brain cancer(9), liver(62), and gall bladder/ biliary tract cancer(62).

Reported diagnostic windows were shortest (i.e. all under six months) for childhood/adolescent bone cancers(65,83), adolescent germ cell tumours(83), oesophageal cancer(62), gastric cancer(62), pancreatic cancer(62), prostate cancer(66,71,77), breast cancer(1,66,71,77,84), malignant melanoma(66,77,78,83), endometrial cancer(62), ovarian cancer(62), and gynaecological cancers combined(66).

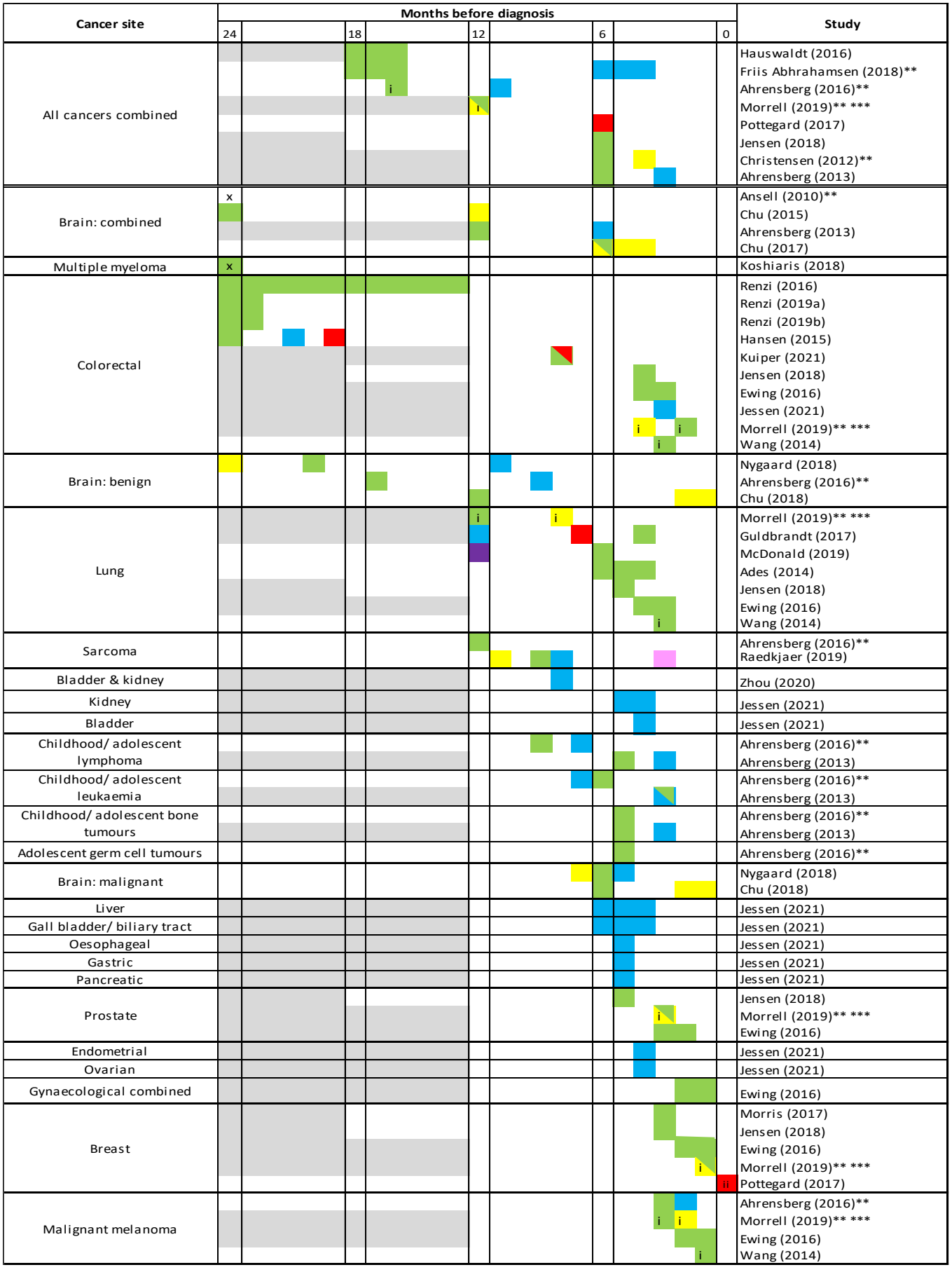

Cancer site
Months before diagnosis
24
18
12
6
0
Study
All cancers combined
Hauswaldt (2016)
Friis Abhrahamsen (2018)**
Ahrensberg (2016)**
Morrell (2019)** ***
Pottegard (2017)
Jensen (2018)
Christensen (2012)**
Ahrensberg (2013)
Brain: combined
Ansell (2010)**
Chu (2015)
Ahrensberg (2013)
Chu (2017)
Multiple myeloma
Koshiaris (2018)
Colorectal
Renzi (2016)
Renzi (2019a)
Renzi (2019b)
Hansen (2015)
Kuiper (2021)
Jensen (2018)
Ewing (2016)
Jessen (2021)
Morrell (2019)** ***
Wang (2014)
Brain: benign
Nygaard (2018)
Ahrensberg (2016)**
Chu (2018)
Lung
Morrell (2019)** ***
Guldbrandt (2017)
McDonald (2019)
Ades (2014)
Jensen (2018)
Ewing (2016)
Wang (2014)
Sarcoma
Ahrensberg (2016)**
Raedkjaer (2019)
Bladder & kidney
Zhou (2020)
Kidney
Jessen (2021)
Bladder
Jessen (2021)
Childhood/ adolescent lymphoma
Ahrensberg (2016)**
Ahrensberg (2013)
Childhood/ adolescent leukaemia
Ahrensberg (2016)**
Ahrensberg (2013)
Childhood/ adolescent bone tumours
Ahrensberg (2016)**
Ahrensberg (2013)
Adolescent germ cell tumours
Ahrensberg (2016)**
Brain: malignant
Nygaard (2018)
Chu (2018)
Liver
Jessen (2021)
Gall bladder/ biliary tract
Jessen (2021)
Oesophageal
Jessen (2021)
Gastric
Jessen (2021)
Pancreatic
Jessen (2021)
Prostate
Jensen (2018)
Morrell (2019)** ***
Ewing (2016)
Endometrial
Jessen (2021)
Ovarian
Jessen (2021)
Gynaecological combined
Ewing (2016)
Breast
Morris (2017)
Jensen (2018)
Ewing (2016)
Morrell (2019)** ***
Pottegard (2017)
Malignant melanoma
Ahrensberg (2016)**
Morrell (2019)** ***
Ewing (2016)
Wang (2014)
Consultations in primary care
Consultations in secondary care
Diagnostic test use
Diagnostic test findings
Prescriptions
Surgical procedures

***Figure 2.3. Longest diagnostic window* for patients diagnosed with each cancer, by study and event type, ranked by diagnostic window length***

**The earliest point in time before diagnosis when a change was observed in a relevant clinical event type. Where multiple figures were given by a study for an event type or patient groups, the earliest single figure is shown. Therefore, the figure shown may only apply to specific groups of patients with that cancer. For studies using longer/ shorter time intervals than months (e.g. quarters, days), the equivalent range of months are highlighted. White space indicates the lookback period before diagnosis used by the study. Where no grey shading is shown, the study lookback period exceeded two years. **Study included two different methods yielding different results; the results of primary focus in the study abstract/ conclusions are shown here. ***Study examined 'GP' and 'specialist' consultations; these were assigned to primary and secondary care consultations, respectively. i Estimated by literature review authors using graphs or tables provided. ii No change before diagnosis. X Longest diagnostic window exceeded two years and is shown in Table 2.*

***Table 2.2. Study settings, cohorts, and longest maximum diagnostic window* for patients diagnosed with each cancer, by study and event type, ranked by diagnostic window length***

Cancers are ranked by maximum diagnostic window length, and within each cancer, studies are ranked by maximum diagnostic window length

*The earliest point in time before diagnosis when a change was observed in a relevant clinical event type. Where multiple figures were given by a study for an event type or patient groups, the earliest single figure is shown. Therefore, the figure shown may only apply to specific groups of patients with that cancer. For studies using longer/ shorter time intervals than months (e.g. quarters, days), the equivalent range of months are highlighted. **Estimated by literature review authors using graphs or tables provided. ***Study included two different methods yielding different results; the results of primary focus in the study abstract/ conclusions are shown here. ****Sample size was not given for the specific cancer site.

| Study | Setting & cohort (country: context; diagnosis dates; age range (N)) | Methods summary (design; method to identify inflection point; observation period pre-diagnosis) | Longest maximum diagnostic window for a single patient group, for each event type studied |
|---|---|---|---|
| | | **All cancers combined** | |
| Hauswaldt (2016) | Germany: primary care; cancers diagnosed 1996-2006 and with 1+ primary care contact < 18 months pre-diagnosis; age range not stated (N=3,310) | Statistical identification (case-control); earliest quarter when the inter-contact interval (time lag between two consecutive consultations) was shorter among cases than controls; < 18 months pre-diagnosis | 16-18 months (decrease in interval between primary care consultations for any reason) |
| Friis Abhrahamsen (2018) | Denmark: primary care; cancers diagnosed 2008-2015; children aged < 15 years (N=1,386) | Statistical identification (case-control); earliest quarter when rates were significantly higher among cases than controls; < 24 months pre-diagnosis** | 16-18 months (increase in primary care consultations for any reason)<br>4-6 months (increase in primary care diagnostic test use (urine tests, blood tests, pulmonary function, electrocardiography, streptococcal throat infection)) |
| Ahrensberg (2016) | Denmark: primary care; cancers diagnosed 2002-2011; young adults 15-39 years (N=12,306) | Statistical identification (case-control); earliest month when rates were significantly higher among cases than controls (and increase was sustained); < 24 months pre-diagnosis** | 16 months (increase in primary care consultations for any reason)**<br>11 months (increase in primary care blood test use) |
| Morrell (2019) | New South Wales, Australia: primary and secondary care; cancers diagnosed 2006-2015; adults aged 45 years + (N=16,750) | Statistical identification (case-control); earliest month when proportions were significantly higher among cases than controls; < 12 months pre-diagnosis** | 12 months (increase in GP consultations for any reason)<br>12 months (increase in specialist consultations for any reason, emergency day visits and emergency inpatient admissions) |
| Pottegard (2017) | Denmark: primary care; cancers diagnosed 2000-2012; adults (N=353,087) | Visual identification (case-control); earliest month when rates among cases appeared to increase; < 24 months pre-diagnosis | 6 months (increase in new first-time prescription use for any drug) |
| Jensen (2018) | Denmark: primary care; cancers diagnosed 2009-2013; adults aged 50-90 years (N=123,943) | Statistical identification (case-only); earliest month when rates were significantly higher among cases compared to cases in the preceding month; < 18 months pre-diagnosis | 6 months (increase in primary care consultations for any reason for males) |
| Christensen (2012) | Denmark:primary and secondary care; cancers diagnosed 2001-2006; adults aged 40 years + (N=127,210) | Visual identification (case-control); earliest month when rates among cases appeared to increase/ earliest quarter when rates were significantly higher among cases than controls; < 12 months pre-diagnosis** | 6 months (increase in primary care consultations for any reason)<br>4 months (increase in secondary care admissions and outpatient visits for any reason)<br>4 months (increase in diagnostic test use (x ray, ultrasound, endoscopy, biopsies, CAT scan, MRI |

| | | | |
|---|---|---|---|
| | | | scan, angiography) within particular specialties)) |
| Ahrensberg (2013) | Denmark: primary care; cancers diagnosed 2002-2008; children < 16 years (N=1,278) | Statistical identification (case-control); earliest month when rates were significantly higher among cases than controls; < 12 months pre-diagnosis | 6 months (increase in primary care daytime consultations for any reason)<br>3 months (increase in primary care diagnostic test use) |
| | | **Brain: benign & malignant combined** | |
| Ansell (2010) | UK: primary care; cancers diagnosed 1992-1996 ; children 1-14 years (N=195) | Statistical identification (case-control); earliest 6-month interval when rates were significantly higher among cases than controls; < 4 years pre-diagnosis** | 4 years (43-48 months) (increase in primary care consultations (relevant symptoms only), increase in records of relevant symptoms in primary care) |
| Chu (2015) | England; primary and secondary care; cancers diagnosed 1989-2006; children & young adults (N=9,799) | Visual identification (case-only); earliest month when smoothed rates among cases appeared to increase; < 36 months pre-diagnosis | 24 months (increase in primary care consultations for headache and growth/ endocrine disorders)<br>12 months (increase in secondary care consultations for convulsions) |
| Ahrensberg (2013) | Denmark: primary care; cancers diagnosed 2002-2008; children < 16 years (N=298) | Statistical identification (case-control); earliest month when rates were significantly higher among cases than controls; < 12 months pre-diagnosis | 12 months (increase in primary care daytime consultations for any reason)<br>6 months (increase in primary care diagnostic test use) |
| Chu (2017) | England; primary and secondary care; cancers diagnosed 1989-2006; children & young adults (N=9,799) | Visual identification (case-only); earliest month when smoothed rates among cases appeared to increase; < 36 months pre-diagnosis | 6 months (increase in primary care consultations for relevant symptoms only)<br>3-6 months (increase in secondary care consultations for relevant symptoms only) |
| | | **Multiple myeloma** | |
| Koshiaris (2018) | England: primary care; cancers diagnosed 2000-2009; adults > 40 years old (N=2,703) | Statistical identification (case-control); earliest quarter when rates were significantly higher/ mean test values were significantly different among cases compared to controls; < 5 years pre-diagnosis | 24 months (increase in primary care consultations for specific symptom groups (back pain, rib pain, chest infections, chest pain, nosebleed))<br>36 months (decrease in primary care mean haemoglobin values) |
| | | **Colorectal** | |
| Renzi (2016) | England: primary care; cancers diagnosed 2005-2006; adults 25 years + (N=1,606) | Statistical identification (case-only); earliest time interval (time interval size varied) when rates were significantly higher among cases compared to cases in the preceding time interval; < 5 years pre-diagnosis | 13-24 months (increase in primary care consultations for any reason) |
| Renzi (2019a) | England: primary care; cancers diagnosed 2005-2010; adults 18 years + (colon only, N=5,745) | Visual identification (case-only); earliest two-month interval when rates among cases appeared to increase; < 5 years pre-diagnosis | 23-24 months (increase in primary care consultations for relevant symptoms for female emergency presenters with 'serious' non gastro-intestinal comorbidities diagnosed/ treated in secondary care) |
| Renzi (2019b) | England: primary care; cancers diagnosed 2005-2010; adults > 18 years (colon only, N=5,745) | Visual identification (case-only); earliest two-month interval when rates among cases appeared to increase; < 5 years pre-diagnosis | 23-24 months (increase in primary care consultations for relevant symptoms for females with proximal colon cancer diagnosed as an emergency) |
| Hansen (2015) | Denmark: primary care; cancers diagnosed 2004-2010; adults aged 40-80 years (N=19,209) | Statistical identification (case-control); earliest month when rates were significantly higher among cases than controls; < 12 months pre-diagnosis (extended to 24 months for some events) | 24 months (increase in primary care consultations for any reason for females with proximal colon cancer)<br>21 months (increase in primary care haemoglobin test use for males with proximal colon cancer)<br>19 months (increase in primary care |

| | | | |
|---|---|---|---|
| | | | haemorrhoid prescription use for females with rectal cancer) |
| Kuiper (2021) | Netherlands: primary care; cancers diagnosed 2007-2014; age range not stated (N=6,087) | Statistical identification (case-control); earliest month when rates were significantly higher among cases than controls; < 12 months pre-diagnosis | 8 months (increase in primary care consultations for any reason for patients with proximal colon cancer)<br>8 months (increase in prescriptions for any drug for patients with proximal colon cancer) |
| Jensen (2018) | Denmark: primary care; cancers diagnosed 2009-2013; adults aged 50-90 years (N=17,138) | Statistical identification (case-only); earliest month when rates were significantly higher among cases compared to cases in the preceding month; < 18 months pre-diagnosis | 4 months (increase in primary care consultations for any reason for males who usually consult 'rarely') |
| Ewing (2016) | Sweden: primary care; cancers diagnosed 2011 and with 1+ primary care contact < 12 months pre-diagnosis; adults (N=753) | Visual identification (case-control); earliest week (reported as days) when rates among cases appeared to increase; < 12 months pre-diagnosis | 3-4 months (100 days) (increase in primary care consultations for any reason, increase in records of diagnostic codes in primary care) |
| Jessen (2021) | Denmark: primary and secondary care; cancers diagnosed 2014-2018; age range not stated (colon N=15,017, rectal N=7,176) | Statistical identification (case-only); earliest month when rates were significantly higher among cases compared to cases in the preceding month; < 12 months pre-diagnosis | 3 months (increase in colonoscopy use for patients with colon or rectal cancer) |
| Morrell (2019) | New South Wales, Australia: primary and secondary care; cancers diagnosed 2006-2015; adults aged 45 years + (N=2,077) | Statistical identification (case-control); earliest month when proportions were significantly higher among cases than controls; < 12 months pre-diagnosis** | 2 months (increase in GP consultations for any reason)**<br>4 months (increase in emergency inpatient admissions)** |
| Wang (2014) | UK; primary care; cancers diagnosed 1997-2006; adults (N=12,189) | Statistical identification (case-only); earliest month when rates were significantly higher among cases compared to cases in the preceding month (identified by literature review authors); < 24 months pre-diagnosis | 3 months (increase in primary care consultations for any reason for males)** |
| | | **Brain: benign** | |
| Nygaard (2018) | Denmark; primary and secondary care; cancers diagnosed 2009-2014; adults (N=3,654) | Statistical identification (case-control); earliest month when rates were significantly higher among cases than controls; < 24 months pre-diagnosis | 20 months (increase in primary care consultations for any reason, for females)<br>24 months (increase in secondary care consultations in Ear-Nose-Throat speciality for males/ females, all other hospital contacts for males)<br>11 months (increase in radiology test use for males) |
| Ahrensberg (2016) | Denmark: primary care; cancers diagnosed 2002-2011; young adults 15-39 years N=1,569) | Statistical identification (case-control); earliest month when rates were significantly higher among cases than controls (and increase was sustained); < 24 months pre-diagnosis** | 17 months (increase in primary care consultations for any reason)<br>9 months (increase in primary care blood test use)<br>6 months (increase in primary care psychometric test use) |
| Chu (2018) | England; primary and secondary care; cancers diagnosed 1989-2006; children & young adults (N=9,799)**** | Visual identification (case-only); earliest month when smoothed rates among cases appeared to increase; < 36 months pre-diagnosis | 12 months (increase in primary care consultations for relevant symptoms only)<br>1-2 months (increase in secondary care consultations for relevant symptoms only) |
| | | **Lung** | |

| | | | |
|---|---|---|---|
| Morrell (2019) | New South Wales, Australia: primary and secondary care; cancers diagnosed 2006-2015; adults aged 45 years + (N=1,235) | Statistical identification (case-control); earliest month when proportions were significantly higher among cases than controls; < 12 months pre-diagnosis** | 12 months (increase in GP consultations for any reason)**<br>8 months (increase in specialist consultations for any reason)** |
| Guldbrandt (2017) | Denmark: primary care; cancers diagnosed 2003-2012; adults aged 40-90 years (N=34,017) | Statistical identification (case-control); earliest month when rates were significantly higher among cases than controls; < 12 months pre-diagnosis | 4 months (increase in primary care consultations for any reason)<br>12 months (increase in first-time primary care lung function test use)<br>12 months (increase in first-time radiology test use)<br>7 months (increase in new COPD prescription use) |
| McDonald (2019) | UK: primary care; cancers diagnosed 1887-2018; adults (N=26,379) | Statistical identification (case-control); earliest two-month interval when rates/ proportions were significantly higher among cases than controls; < 24 months pre-diagnosis | 6 months (increase in records of relevant symptoms in primary care)<br>12 months (increase in proportion of patients in primary care with high CRP test values) |
| Ades (2014) | Devon, UK: primary care; cancers diagnosed 1998-2002; age range not stated (N=247) | Visual identification (case-control); earliest quarter when rates among cases appeared to increase; < 24 months pre-diagnosis | 4-6 months (increase in records of two relevant symptoms per quarter in primary care) |
| Jensen (2018) | Denmark: primary care; cancers diagnosed 2009-2013; adults aged 50-90 years (N=17,861) | Statistical identification (case-only); earliest month when rates were significantly higher among cases compared to cases in the preceding month; < 18 months pre-diagnosis | 5 months (increase in primary care consultations for any reason for females who usually consult with 'average' frequency) |
| Ewing (2016) | Sweden: primary care; cancers diagnosed 2011 and with 1+ primary care contact < 12 months pre-diagnosis; adults (N=373) | Visual identification (case-control); earliest week (reported as days) when rates among cases appeared to increase; < 12 months pre-diagnosis | 3-4 months (100 days) (increase in primary care consultations for any reason, increase in records of diagnostic codes in primary care) |
| Wang (2014) | UK; primary care; cancers diagnosed 1997-2006; adults (N= 11,081) | Statistical identification (case-only); earliest month when rates were significantly higher among cases compared to cases in the preceding month (identified by literature review authors); < 24 months pre-diagnosis | 3 months (increase in primary care consultations for any reason for males/ females)** |
| | | **Sarcoma** | |
| Ahrensberg (2016) | Denmark: primary care; cancers diagnosed 2002-2011; young adults 15-39 years (soft tissue sarcoma only N=315) | Statistical identification (case-control); earliest month when rates were significantly higher among cases than controls (and increase was sustained); < 24 months pre-diagnosis** | 12 months (increase in primary care consultations for any reason) |
| Raedkjaer (2019) | Denmark: primary and secondary care; cancers diagnosed 2000-2013; adults (N=2,167) | Statistical identification (case-control); earliest month when rates were significantly higher among cases than controls; < 24 months pre-diagnosis | 9 months (increase in primary care consultations for any reason)<br>11 months (increase in secondary care inpatient consultations for any reason, within orthopaedic surgery, dermatology, plastic surgery)<br>3 months (increase in secondary care surgery, within orthopaedic surgery, dermatology, plastic surgery)<br>8 months (increase in secondary care paraclinical examinations) |
| | | **Bladder & kidney combined** | |

| | | | |
|---|---|---|---|
| Zhou (2020) | England; primary and secondary care; cancers diagnosed 2012-2015; adults 25 years+ (N=2,971) | Statistical identification through model comparison (case-only); models of monthly rates were fitted with different likely inflection points, with the model with optimal goodness of fit chosen; < 12 months pre-diagnosis | 8 months (increase in x-ray use) |
| | | **Kidney** | |
| Jessen (2021) | Denmark: primary and secondary care; cancers diagnosed 2014-2018; age range not stated (N=4,224) | Statistical identification (case-only); earliest month when rates were significantly higher among cases compared to cases in the preceding month; < 12 months pre-diagnosis | 4-5 months (increase in abdominal CT use) |
| | | **Bladder** | |
| Jessen (2021) | Denmark: primary and secondary care; cancers diagnosed 2014-2018; age range not stated (N=3,801) | Statistical identification (case-only); earliest month when rates were significantly higher among cases compared to cases in the preceding month; < 12 months pre-diagnosis | 4 months (increase in transvaginal ultrasound use) |
| | | **Childhood/ adolescent lymphoma** | |
| Ahrensberg (2016) | Denmark: primary care; cancers diagnosed 2002-2011; young adults 15-39 years (N=765) | Statistical identification (case-control); earliest month when rates were significantly higher among cases than controls (and increase was sustained); < 24 months pre-diagnosis** | 9 months (increase in primary care consultations for any reason)<br>7 months (increase in primary care blood test use for leukaemia & lymphoma combined) |
| Ahrensberg (2013) | Denmark: primary care; cancers diagnosed 2002-2008; children < 16 years (N=105) | Statistical identification (case-control); earliest month when rates were significantly higher among cases than controls; < 12 months pre-diagnosis | 5 months (increase in primary care daytime consultations for any reason)<br>3 months (increase in use of primary care diagnostic tests) |
| | | **Childhood/ adolescent leukaemia** | |
| Ahrensberg (2016) | Denmark: primary care; cancers diagnosed 2002-2011; young adults 15-39 years (N=386) | Statistical identification (case-control); earliest month when rates were significantly higher among cases than controls (and increase was sustained); < 24 months pre-diagnosis** | 6 months (increase in primary care consultations for any reason)<br>7 months (increase in primary care blood test use for leukaemia & lymphoma combined) |
| Ahrensberg (2013) | Denmark: primary care; cancers diagnosed 2002-2008; children < 16 years (N=354) | Statistical identification (case-control); earliest month when rates were significantly higher among cases than controls; < 12 months pre-diagnosis | 3 months (increase in primary care daytime consultations for any reason)<br>3 months (increase in use of primary care diagnostic tests) |
| | | **Childhood/ adolescent bone tumours** | |
| Ahrensberg (2016) | Denmark: primary care; cancers diagnosed 2002-2011; young adults 15-39 years (N=144) | Statistical identification (case-control); earliest month when rates were significantly higher among cases than controls (and increase was sustained); < 24 months pre-diagnosis** | 5 months (increase in primary care consultations for any reason) |
| Ahrensberg (2013) | Denmark: primary care; cancers diagnosed 2002-2008; children < 16 years (N=65) | Statistical identification (case-control); earliest month when rates were significantly higher among cases than controls; < 12 months pre-diagnosis | 5 months (increase in primary care daytime consultations for any reason)<br>3 months (increase in use of primary care diagnostic tests) |
| | | **Adolescent germ cell tumours** | |
| Ahrensberg (2016) | Denmark: primary care; cancers diagnosed 2002-2011; young adults 15-39 years (N=1,837) | Statistical identification (case-control); earliest month when rates were significantly higher among cases than controls (and increase was sustained); < 24 months pre-diagnosis** | 5 months (increase in primary care consultations for any reason) |

| | | | |
|---|---|---|---|
| | | **Brain: malignant** | |
| Nygaard (2018) | Denmark; primary and secondary care; cancers diagnosed 2009-2014; adults (N=2,272) | Statistical identification (case-control); earliest month when rates were significantly higher among cases than controls; < 24 months pre-diagnosis | 6 months (increase in primary care consultations for any reason for males)<br>7 months (increase in secondary care consultations in neurology, for females)<br>5 months (increase in radiology test use for males/ females) |
| Chu (2018) | England; primary and secondary care; cancers diagnosed 1989-2006; children & young adults (N=9,799)**** | Visual identification (case-only); earliest month when smoothed rates among cases appeared to increase; < 36 months pre-diagnosis | 6 months (increase in primary care consultations for relevant symptoms only)<br>1-2 months (increase in secondary care consultations for relevant symptoms only) |
| | | **Liver** | |
| Jessen (2021) | Denmark: primary and secondary care; cancers diagnosed 2014-2018; age range not stated (N= 2,028) | Statistical identification (case-only); earliest month when rates were significantly higher among cases compared to cases in the preceding month; < 12 months pre-diagnosis | 4-6 months (increase in abdominal ultrasound use) |
| | | **Gall bladder/ biliary tract** | |
| Jessen (2021) | Denmark: primary and secondary care; cancers diagnosed 2014-2018; age range not stated (N=906) | Statistical identification (case-only); earliest month when rates were significantly higher among cases compared to cases in the preceding month; < 12 months pre-diagnosis | 4-6 months (increase in abdominal ultrasound use) |
| | | **Oesophageal** | |
| Jessen (2021) | Denmark: primary and secondary care; cancers diagnosed 2014-2018; age range not stated (N= 2,263) | Statistical identification (case-only); earliest month when rates were significantly higher among cases compared to cases in the preceding month; < 12 months pre-diagnosis | 5 months (increase in gastroscopy use) |
| | | **Gastric** | |
| Jessen (2021) | Denmark: primary and secondary care; cancers diagnosed 2014-2018; age range not stated (N=2,660) | Statistical identification (case-only); earliest month when rates were significantly higher among cases compared to cases in the preceding month; < 12 months pre-diagnosis | 5 months (increase in gastroscopy use) |
| | | **Pancreatic** | |
| Jessen (2021) | Denmark: primary and secondary care; cancers diagnosed 2014-2018; age range not stated (N=4,304) | Statistical identification (case-only); earliest month when rates were significantly higher among cases compared to cases in the preceding month; < 12 months pre-diagnosis | 5 months (increase in gastroscopy or endoscopic retrograde cholangiopancreatography use) |
| | | **Prostate** | |
| Jensen (2018) | Denmark: primary care; cancers diagnosed 2009-2013; adults aged 50-90 years (N=19,348) | Statistical identification (case-only); earliest month when rates were significantly higher among cases compared to cases in the preceding month; < 18 months pre-diagnosis | 5 months (increase in primary care consultations for any reason for males who usually consult with 'average' frequency) |
| Morrell (2019) | New South Wales, Australia: primary and secondary care; cancers diagnosed 2006-2015; adults aged 45 years + (N=3,960) | Statistical identification (case-control); earliest month when proportions were significantly higher among cases than controls; < 12 months pre-diagnosis** | 3 months (increase in GP consultations for any reason)**<br>3 months (increase in specialist consultations for any reason and emergency day visits)** |
| Ewing (2016) | Sweden: primary care; cancers diagnosed 2011 and with 1+ primary care contact < 12 | Visual identification (case-control); earliest week (reported as days) when rates among cases appeared to increase; < 12 months pre-diagnosis | 2-3 months (80 days) (increase in primary care consultations for any reason, increase in records of diagnostic codes in primary care) |

| | | | |
|---|---|---|---|
| | months pre-diagnosis; adults (N=1,257) | | |
| | | **Endometrial** | |
| Jessen (2021) | Denmark: primary and secondary care; cancers diagnosed 2014-2018; age range not stated (N=3,517) | Statistical identification (case-only); earliest month when rates were significantly higher among cases compared to cases in the preceding month; < 12 months pre-diagnosis | 4 months (increase in transvaginal ultrasound use) |
| | | **Ovarian** | |
| Jessen (2021) | Denmark: primary and secondary care; cancers diagnosed 2014-2018; age range not stated (N=2,002) | Statistical identification (case-only); earliest month when rates were significantly higher among cases compared to cases in the preceding month; < 12 months pre-diagnosis | 4 months (increase in transvaginal ultrasound use) |
| | | **Gynaecological combined** | |
| Ewing (2016) | Sweden: primary care; cancers diagnosed 2011 and with 1+ primary care contact < 12 months pre-diagnosis; adults (N=327) | Visual identification (case-control); earliest week (reported as days) when rates among cases appeared to increase; < 12 months pre-diagnosis | 1-2 months (50 days) (increase in primary care consultations for any reason, increase in records of diagnostic codes in primary care) |
| | | **Breast** | |
| Morris (2017) | West Midlands, England: primary care; cancers diagnosed 1989-2006; adults 50-70 years (N=786) | Visual identification (case-only); earliest month when rates among cases appeared to increase; < 18 months pre-diagnosis | 3 months (increase in primary care consultations for breast-related symptoms) |
| Jensen (2018) | Denmark: primary care; cancers diagnosed 2009-2013; adults aged 50-90 years (N=18,396) | Statistical identification (case-only); earliest month when rates were significantly higher among cases compared to cases in the preceding month; < 18 months pre-diagnosis | 3 months (increase in primary care consultations for any reason) |
| Ewing (2016) | Sweden: primary care; cancers diagnosed 2011 and with 1+ primary care contact < 12 months pre-diagnosis; adults (N=947) | Visual identification (case-control); earliest week (reported as days) when rates among cases appeared to increase; < 12 months pre-diagnosis | 1-2 months (50 days) (increase in primary care consultations for any reason, increase in records of diagnostic codes in primary care) |
| Morrell (2019) | New South Wales, Australia: primary and secondary care; cancers diagnosed 2006-2015; adults aged 45 years + (N=1,999) | Statistical identification (case-control); earliest month when proportions were significantly higher among cases than controls; < 12 months pre-diagnosis** | 1 month (increase in GP consultations for any reason)**<br>1 month (increase in specialist consultations for any reason)** |
| Pottegard (2017) | Denmark: primary care; cancers diagnosed 2000-2012; adults (N=51,774) | Visual identification (case-control); earliest month when rates among cases appeared to increase; < 24 months pre-diagnosis | No increase pre-diagnosis (new first-time prescriptions use for any drug) |
| | | **Malignant melanoma** | |
| Ahrensberg (2016) | Denmark: primary care; cancers diagnosed 2002-2011; young adults 15-39 years (N=2,501) | Statistical identification (case-control); earliest month when rates were significantly higher among cases than controls (and increase was sustained); < 24 months pre-diagnosis** | 3 months (increase in primary care consultations for any reason)<br>2 months (increase in primary care blood test use) |
| Morrell (2019) | New South Wales, Australia: primary and secondary care; cancers diagnosed 2006-2015; adults aged 45 years + (N=2,070) | Statistical identification (case-control); earliest month when proportions were significantly higher among cases than controls; < 12 months pre-diagnosis** | 3 months (increase in GP consultations for any reason)**<br>2 months (increase in specialist consultations for any reason)** |

| | | | |
|---|---|---|---|
| Ewing (2016) | Sweden: primary care; cancers diagnosed 2011 and with 1+ primary care contact < 12 months pre-diagnosis; adults (N=459) | Visual identification (case-control); earliest week (reported as days) when rates among cases appeared to increase; < 12 months pre-diagnosis | 1-2 months (60 days) (increase in primary care consultations for any reason, increase in records of diagnostic codes in primary care) |
| Wang (2014) | UK; primary care; cancers diagnosed 1997-2006; adults (N=4,352) | Statistical identification (case-only); earliest month when rates were significantly higher among cases compared to cases in the preceding month (identified by literature review authors); < 24 months pre-diagnosis | 1 month (increase in primary care consultations for any reason)** |

### 2.7.5 Variation by tumour characteristics

Six studies examined variation by tumour characteristics. As indicated by changes in primary care consultation rates in two studies, reported diagnostic windows were generally longer for proximal colon compared to distal colon or rectal cancer(61,69). Increases in prescription rates for any newly-prescribed drug occurred earlier for proximal colon compared to distal colon or rectal cancer,(61) whereas increases in haemorrhoid prescription rates were earlier for rectal compared to colon cancers(8). For brain tumours, window lengths varied by anatomic subsite (e.g. the supratentorial compartment, the midline, or cranial nerves)(73), and for some presenting symptoms (e.g. headache and convulsions), although patterns were complex(76). For lung cancer, diagnostic windows did not vary by stage at diagnosis(81).

### 2.7.6 Variation by patient group

A study examining multiple cancer sites, and one studying all cancers combined found no differences in the length of the potential diagnostic window by sex(75,78), while another including patients with brain cancer commented that differences existed, without specifying the pattern(9). Five other studies did not comment specifically on differences in the diagnostic window length, but did stratify findings by sex(8,10,68,70,71,83). Some noted that sex stratification was needed given gender differences in baseline healthcare utilisation or comorbidities(8,9,70,71). Where examined, there was little evidence of variation in the inflection point by patients' usual/background consultation frequencies(71).

A study reported no differences in the length of the potential diagnostic window between patients diagnosed with colorectal cancer who were diagnosed either through an emergency presentation or through other diagnostic routes(79). Two others examining colon cancer found likely differences in the diagnostic window length when considering comorbidity status and diagnostic route(69,70). For example, women with 'serious' non gastro-intestinal comorbidities who were diagnosed with colon cancer as an emergency had longer diagnostic windows, compared to non-comorbid women diagnosed either through emergency or non-emergency routes(70).

## 2.8 Discussion

### 2.8.1 Key findings

Evidence from electronic patient records indicates that for 15 common cancers, some patients begin to present at least six months before diagnosis. In the case of colorectal, brain tumours, and multiple myeloma, some studies suggest this may be even sooner. The majority of this evidence was produced by studies examining increases in primary care consultations (including consultations for any reason or for specific presenting symptoms), but also included studies examining increases in secondary care consultations, diagnostic test use or changes in diagnostic test findings.

Longer diagnostic windows were identified for specific brain and colon cancer sub-sites, and for brain cancer patients, as indicated by increases in consultations for specific symptoms. Where studied, there was no evidence of, or limited variability in diagnostic window length by stage at diagnosis, sex, usual consultation frequency, or emergency presentation status (except for women with multi-morbidities diagnosed with colon cancer, and women diagnosed with proximal colon cancer)(69–71,78,79,81).

### 2.8.2 Comparison with existing literature

I am not aware of previous reviews examining the length of potential diagnostic windows in patients with cancer. Some previous studies have estimated diagnostic intervals for individual patients, for example, from a presentation that is deemed *a priori* to be the first relevant one to the time of subsequently diagnosed cancer(13–15). These studies rely on assumptions about how to define the 'first relevant' presentation, and achieving consistent definitions between studies is challenging, particularly in patients with morbidity who regularly consult for unrelated reasons(86). The reviewed studies use a population approach in order to identify the earliest point at which healthcare utilisation rates change in some patients, avoiding the need for any such assumptions(72).

### 2.8.3 Limitations of the reviewed evidence

There are several limitations of the reviewed evidence. Firstly, evidence for 13 cancer sites (e.g. pancreatic cancer) was limited to single studies. It should be noted that for some of the cancer sites with the longest diagnostic windows (multiple myeloma and brain tumours), evidence of particularly long diagnostic windows of over one year was limited to one or two studies each. In addition, although the reviewed studies have the potential to illuminate disparities in the length of the diagnostic window between different patient groups, this has not yet been examined with regard to ethnicity, comorbidities, and age.

Studies used different methods to identify the onset of changes in healthcare utilisation (i.e. the timing of inflection points), potentially because they considered the measurement of diagnostic window length as a secondary or subsidiary aim. The exact timing of inflection points seems sensitive to the type of comparison used (i.e. whether through visual inspection or statistical approaches) and the study type used (i.e. case-only or case-control), as illustrated by studies that

used more than one approach(75,77,80,82,83). I have summarised and reflected on these methodological issues and related recommendations in Table 2.1.

In principle, the length of observed diagnostic windows may be influenced by the rate of tests performed, or the completeness of recording (e.g. of presenting symptoms). Regarding testing, greater or lower use of tests by doctors (e.g. as can be encountered in different study eras or different health systems) could impact the background rate of abnormal test results in either cases or controls in a population(10,87). If the background rate of testing is higher in cases, diagnostic windows may be longer in case-control studies. Regarding consultations, their occurrence is recorded reliably in electronic health record patient systems, so background rates should not differ systematically between cases and controls. However, the recording of a specific presenting symptom during a consultation could be mediated by the doctor's perception of the patient's risk of serious disease(88,89). Therefore, diagnostic windows related to rates of specific symptom presentations could be subject to similar biases to the recording of abnormal test results.

Power to detect inflection points is driven by the number of events in given time periods. Therefore, power may have been limited in certain studies, for example those using small samples of patients, short time units of analysis (e.g. monthly rather than quarterly rates), or examining relatively rare healthcare utilisation event types. However, shorter time units potentially offer more precise estimates of the timing of inflection points (for example, identifying the month, rather than the quarter where healthcare use begins to change from baseline). Limited power may also result from appropriate stratification by cancer site or gender.

All studies identifying differences in the inflection point between patient or tumour groups did so using stratified analyses. (78) This approach is problematic for two reasons. Firstly, there was no formal test for significant differences in the timing of the inflection point between groups. If, for example, a study claimed that pre-diagnostic consultation rates began to rise five months before diagnosis for men, and four months for women, the degree of certainty with which it can be stated that the inflection point was earlier for men remains unclear unless formal comparisons are performed. Secondly, in case-only studies that directly compared the point that healthcare use changed in one group against the point that it changed in another, there was generally no adjustment for confounding, i.e. other characteristics of these groups that could possibly explain the difference (e.g. Wang et al (78)). Hypothetically, if healthcare use increased earlier for patients with a higher comorbidity score, and patients with more comorbidities were generally older, then the observed differences between patients with high and low comorbidity scores may reflect longer diagnostic windows associated with increasing age, rather than comorbidity per se(90). Finally, the length of the observation period before diagnosis varied by study. Longer observation periods to capture changing healthcare use should be recommended, as maximum reported diagnostic window lengths for some cancer sites are as long as two or three years.

### 2.8.4 Limitations of the review

A limitation of the review is the use of a single database (PubMed) to identify articles; other databases such as MedLine were not included. In addition, it is possible that I did not identify some relevant papers, as some studies evaluated the diagnostic window to fulfil a subsidiary aim, so it may not have been mentioned in the abstract, title, or keywords. I therefore maximised coverage by

including articles obtained via expert recommendation and searching the reference lists of articles already included.

Some of the observed variation between cancer sites in this review are likely explained by the aforementioned methodological variation between studies. Therefore, I have presented diagnostic windows by both cancer site and study in Figure 2.3. As an illustration, in keeping with other case-only studies, I identified the inflection point in figures provided by Wang et al as the first month when estimates among cases were different to the month before (their confidence intervals did not overlap). These figures would vary considerably if identified as the first month when estimates were different to the start of the observation period, however, this approach could be affected by gradual increases in healthcare utilisation as patients aged over the course of the study. Furthermore, due to stratification (e.g. by gender), the diagnostic window I extracted for some cancer sites may apply to a subset of patients, rather than to all patients diagnosed with that cancer (details are available in Table 2.2).

An additional source of variation between cancer sites was heterogeneity between studies in the healthcare events studied. Studies examined different healthcare events (e.g. consultations, prescriptions, tests), and also defined them in different ways. For example, some studies included all primary care consultations, while others only included those for specific symptoms, or via specific contact methods (e.g. face to face). Due to the other sources of variation between studies noted above (e.g. cancer site examined, study design), it was not possible in this review to quantify variation in diagnostic window length according to the type of healthcare event studied and how it was defined. A handful of studies explicitly examined and discussed this issue(8,63,68,75), but further studies are needed, particularly those that include consultations for specific symptoms, and situated in healthcare systems other than Denmark. There is some evidence that diagnostic windows could vary according to the order in which healthcare events tend to occur in the patient's diagnostic pathway. For example, a consultation with a GP tends to be the first event to occur, so studies examining this event may reveal earlier changes compared to those examining changes in diagnostic test use or abnormal test results(75).

### 2.8.5 Implications

The length of the diagnostic window after initial presentation to healthcare services could be influenced by tumour factors (e.g. cancer site, tumour aggressiveness, symptom signature), patient factors (e.g. comorbidities, patient engagement with healthcare services), and healthcare factors (e.g. type, timeliness, and availability of diagnostic investigations, and monitoring ('safety-netting') protocols). Longer diagnostic windows could indicate opportunities to diagnose cancer sooner in some patients. These could arise in patients with cancers characterised by early onset but non-specific symptoms, which are often either not immediately investigated, or investigated with non-specific tests that lead to complex and prolonged diagnostic pathways to eventual diagnosis. However, the exact mechanisms leading to potentially avoidable delays have not been established in the reviewed literature. Further research is therefore needed to help targeting of interventions to support the diagnostic process. In future studies, further stratification of by healthcare event type could help to better understand where the greatest opportunities lie for earlier diagnosis. Many studies in this review examined consultation use in general; however, those that studied more specific healthcare events such as rates of consultations for specific symptoms or changing blood test results provide more information about what kinds of signs and symptoms could potentially be

early indicators of as-yet undiagnosed cancer, and hence considered for inclusion in subsequent risk prediction studies.

Although the literature suggests that time to diagnosis could be shortened in some patients, it may not necessarily reduce the proportion of patients diagnosed at an advanced stage of disease, because slowly progressing tumours may be over-represented among patients who experience long diagnostic intervals(57,81,90). In addition, by its nature, the onset of a diagnostic window identified from a population will reflect the group of patients with the longest intervals between first presentation and diagnosis, with most patients having shorter diagnostic intervals. A more detailed understanding is needed regarding the proportion of patients whose diagnosis could be expedited, and by how long.

The findings indicate that there is potential to harness electronic health records to inform the management of patients in practice. Electronic health records could be used to develop diagnostic "e-triggers"; these could flag patients in whom the suspicion of cancer may require monitoring or repeat assessment(10,63,67,72), for example, if patients consult more frequently than usual(71,79), or receive particular prescriptions e.g. for haemorrhoids(8), which could raise suspicion of particular cancers. This prospect is particularly promising for patients in contact with healthcare services who are at increased underlying cancer risk (for example, due to their age or pre-existing comorbidities), who do not present with any specific 'alarm' symptoms for cancer that would usually qualify them for urgent referral(12).

In practice, identifying increased consultation frequency in individuals in a timely manner may be challenging, because most patients do not consult regularly at baseline(64,71). Where an increase in healthcare use is identified for a patient, the predictive value of increased consultation frequency is still likely to be low if considered in isolation, as consultations are relatively common events in the general population, compared to cancer. Therefore, an observed change in a patient's individual consultation frequency may need to be combined with other clinical features (e.g. by presenting symptom, or history of additional diagnostic investigations) to better inform risk quantification(64). Relatively little is still known about the risk of all cancers combined and the full range of cancer sites for cohorts of patients presenting with non-specific symptoms, including how this risk is modified by patient information (e.g. age, sex) that is readily available to practicing clinicians during the consultation (see Chapter 1). Arguably, research should be prioritised that aims to address this obvious gap in the evidence underpinning diagnostic guidelines for suspected cancer over and above efforts to incorporate aspects of healthcare use history into risk prediction, particularly given the statistical challenges involved.

## 2.9 Chapter summary

Evidence of changing healthcare utilisation before cancer diagnosis recorded in electronic health records can be used to identify tumour or patient groups in which faster diagnosis could be achieved, and how much faster might be possible. The review revealed that with future improvements to the diagnostic process and diagnostic technologies, some patients could potentially be diagnosed with cancer at least six months earlier. Future research is needed to confirm this, especially studies that address the identified methodological issues, and explore variation by tumour and patient groups, and additional cancer sites. It was especially notable that primary care consultation rates were detectable six months before diagnosis for some patients diagnosed with cancers that are often characterised by non-specific ('vague') symptoms (e.g. multiple myeloma, lung cancer, and sarcoma). This realisation, combined with known gaps in available evidence informing UK diagnostic guidelines for patients presenting in primary care with vague symptoms, indicated that primary studies seeking to quantify cancer risk in patients presenting with vague symptoms could help GPs assess cancer risk for individual patients and contribute to early detection efforts, and has provided the motivation for subsequent work.

However, by reflecting on the evidence included in this review, I concluded that a study using changing healthcare use to stratify cancer risk in patients presenting with fatigue would present several methodological challenges. Given the lack of evidence establishing cancer risk in patients with fatigue according to simple information readily available to clinicians in patients' healthcare records (e.g. age, sex, other presenting symptoms etc), the incorporation of healthcare use into my subsequent risk prediction studies could not be prioritised.

# 3. Chapter 3. Methodological issues in studies of undetected disease risk in symptomatic cohorts using electronic health records

## 3.1 Chapter rationale

In the development of the three empirical chapters examining cancer and other disease risk associated with fatigue as a presenting symptom, I addressed a range of methodological issues involved in using electronic health records (EHRs) to examine disease risk using cohort study designs. Some of these issues have been previously described, but some remain underdeveloped in prior relevant literature and required further examination as part of this thesis. In this chapter, I discuss these methodological issues, including a) choosing appropriate comparisons and statistical methods; b) defining symptomatic cohorts including combinations of features; c) defining outcomes; and d) developing EHR phenotypes. I then explain the approaches I used in my empirical studies on fatigue. This learning will help achieve methodological transparency and reproducibility of my studies, and support other researchers to develop future similar studies using EHRs. Full details of methods specific to each study are further available in the respective chapter.

## 3.2 Publication

Some of the methods described in this Chapter relate to previously published work detailed in Chapters 4, 5, and 6.

## 3.3 Supervisory contributions

The development of this chapter was supervised by Dr Matthew Barclay, with additional contributions from Prof Georgios Lyratzopoulos.

## 3.4 Introduction

Assessing risk of disease relating to a potential prodromal sign or symptom ideally requires cohort studies. Historically, this was only possible through resource-intensive studies of large prospective cohorts followed-up for many years, such as the Framingham Heart Study (91). The infeasibility of conducting such cohort studies meant that evidence on the predictive value of symptoms was limited to case-control or case-only analyses investigating recorded symptoms and signs prior to diagnosis in patients with cancer. These used limited datasets such as local or national population-based disease registries (e.g. the CAPER studies, originally based on manual collation of data on patients from few general practices in the UK (Exeter), or Norway (38,92)) or clinician-curated collections of data such as primary care cancer diagnosis audits or other bespoke designs (e.g. NCDA (17,18,30,52)). Since 2006(93), the availability of large-scale primary care electronic health records has increasingly allowed researchers to examine disease risk in large patient cohorts (91,94).

Primary care electronic health records (EHRs) offer many advantages over traditional data sources. They cover large populations with good representativeness, enabling the examination of risk for rarer diseases and within specific patient strata. Data are routinely collected and can be analysed retrospectively, which dramatically reduces the cost and time involved in recruiting such large samples, compared to a bespoke cohort study. Patients do not have to actively provide consent for participation (although they can opt out), reducing selection bias, generating samples approximately representative of the wider population(91).

Yet primary care EHR datasets are not designed for use in research – typically being intended either for billing purposes or direct patient care (95). Using EHR datasets in research presents specific challenges, over and above the challenges endemic to risk studies of symptomatic cohorts. These include overly granular or 'messy' coding systems for the purposes of research; missing or selectively recorded data items; loss to follow up as patients move practices; difficulties accurately determining the patients' precise health-state (e.g. exactly when symptoms start and end, and whether they are experienced concurrently), since their status is only recorded when patients present in primary care and records do not typically include information on symptom duration; and poor generalisability of risk estimates to healthcare settings systems with different diagnostic services organisation and clinical practice guidelines. In this section I set out issues that need to be considered when designing studies to examine disease risk in symptomatic patients using primary care EHR data, followed by exposition of how I addressed these issues when studying disease risk in patients presenting to primary care with fatigue in my studies.

## 3.5 Using primary care EHRs to study undetected disease risk in symptomatic cohorts

In this section, I discuss the methodological decisions involved in using primary care electronic health records (EHRs) to study short-term disease risk in symptomatic patients. I primarily use examples of symptomatic cohort studies, but also studies patients with abnormal primary care test results (44,45,48–50,55,96–111).

### 3.5.1 Defining a study purpose

Primary care EHR-based studies of short-term disease risk in cohorts of patients presenting with symptoms and/or other clinical features can be categorised into three broad themes (acknowledging that some studies incorporate more than one theme as different components):

1) **'Descriptive risk' studies** typically aim to inform diagnostic guideline decision makers and GPs about risk of undetected disease in cohorts of patients presenting with a symptom, to identify subgroups for whom further diagnostic investigation may be necessary. These can support GPs diagnostic decisions for patient cohorts for whom evidence about disease risk is sparse, for example, risk of cancer and other diseases following primary care presentation with new-onset abdominal pain (98), or cancer incidence in patients with a high normal platelet count(105). These studies typically describe the absolute risk of a disease in a cohort of patients presenting with a symptom, but do not assess whether the symptom is associated with the disease; i.e., whether risk is higher in patients with the symptom than those without.

2) **'Diagnostic value' studies** seek to identify whether a feature (symptom or test result) is associated with increased or decreased disease risk(49,96,97,100–102,112). These studies have been used to identify clinical features that could be incorporated into existing diagnostic guidelines, and/or that could be considered in risk prediction models. For example, one study examined the diagnostic value of inflammatory markers for detecting infections, autoimmune disease, and cancer (101,102). These studies differ from 'descriptive' studies as they seek to assess whether the symptom is associated with increased risk, and - to varying degrees – whether the presence of the symptom can explain the association, by controlling for confounders (e.g. age, sex, comorbidities). Hence, study outputs often include adjusted risk ratios of disease in patients with and without a symptom, using symptomatic cohort designs with matched controls or a reference population.

   These studies can also include those that identify other symptoms or features that add predictive value for a disease *within* a previously defined cohort, for example, cancer incidence in patients with weight loss in combination with/without other signs and symptoms (49)

3) **'Risk prediction' models** combine multiple clinical features and demographic characteristics to generate personalised disease risk scores for individuals that can aid GPs in their decision-

making for presenting patients. Some have also informed clinical guidelines such as UK urgent referral guidelines for suspected cancer(44,45), for example, studies of risk of cancer in patients presenting with selected symptoms (QCancer) (44,45)).

Situating studies within these three broad themes can guide key methodological decisions, including choosing appropriate comparisons, statistical methods, and outcomes. In practice, many existing studies combine elements of both 1) and 2). However, understanding the distinction is important as the statistical methods and results needed are different. 'Diagnostic value' studies generally require measures of the difference in risk between symptomatic patients and a comparison group (with matching or adjustment for confounders), whereas 'descriptive risk' studies require clinically-relevant and easy to interpret measures of absolute risk in a cohort that represents the 'average' patient that the doctor sees.

Finally, while 'risk prediction' models are highlighted in this framework to demonstrate the difference between them and 'descriptive risk'/ 'diagnostic value' studies, they are not the focus of my empirical work and so the following discussion is not applicable to them.

### 3.5.2 Choosing appropriate comparisons

*Comparing against baseline risk*

'Descriptive' studies do not compare disease risk in a symptomatic cohort against a comparison group, because they only aim to identify groups of symptomatic patients at high absolute risk of disease, such as risk that crosses a specific referral threshold. They cannot conclude whether the symptom is associated with increased disease risk. For instance, absolute risk of disease could be 8% for a symptomatic cohort, but its diagnostic value would be low if background risk in an appropriately chosen comparison group was 6%, and conversely, its diagnostic value would be far higher if background risk was 1% (Figure 3.1, diseases 1-4).

In contrast, 'diagnostic value' studies compare risk in patients with a symptom to, for example, the expected risk for patients of the same age and sex. In practice, they often need to incorporate a descriptive element, to establish whether a symptom with high predictive value (i.e., a large relative increase in risk) is also clinically important (i.e., the absolute risk in patients with the symptom is high). For example, Nicholson et al presented hazard ratios of cancer risk in patients presenting with unexpected weight loss compared to matched controls presenting without, assessing the strength of the association between cancer and weight loss presentation. They also presented observed absolute risk of cancer (overall and by cancer site), to inform GPs about which patients were at greatest risk, and which cancer sites to suspect first(97) .

To illustrate, a symptom that is associated with a doubling of risk may not push a patient's risk over a meaningful clinical referral threshold (e.g. typically 3% for cancer), if it meant the absolute risk doubled from 0.5% to 1%. In contrast, the same doubling of risk could be very meaningful if it meant that it pushed the absolute risk from 2% to 4% (Figure 3.1, scenarios 5-8).

| Absolute risk in symptomatic patients is consistent, but excess relative risk increases<br>Same clinical importance, but increasing predictive value | | | |
|---|---|---|---|
| Disease | Absolute risk in symptomatic cohort | Absolute risk in comparison cohort | Relative excess risk |
| 1 | 8% | 6% | 33% |
| 2 | 8% | 4% | 100% |
| 3 | 8% | 2% | 300% |
| 4 | 8% | 1% | 700% |

| Absolute risk in symptomatic patients increases, but excess risk is consistent<br>Increasing clinical importance, but same predictive value | | | |
|---|---|---|---|
| Disease | Absolute risk in symptomatic cohort | Absolute risk in comparison cohort | Relative excess risk |
| 5 | 1% | 0.5% | 100% |
| 6 | 2% | 1% | 100% |
| 7 | 4% | 2% | 100% |
| 8 | 10% | 5% | 100% |

*Figure 3.1. Hypothetical scenarios in which the absolute risk and diagnostic value of a symptom can vary*

Establishing the relative increase in risk in diagnostic value research can be achieved through various methods, including:

- **Symptomatic cohort design with matched controls:** Disease risk in the symptomatic cohort is compared to that of primary care-based controls, who are selected by matching on age, sex, presentation during a similar period, and sometimes GP practice.
- **Symptomatic cohort design with reference population:** Disease risk in the symptomatic cohort is compared to that of a reference comparator group external to the studied population (e.g. the general population, or primary care presenters without the symptom), which is usually case-mix adjusted to the symptomatic cohort based on age and sex.
- **Full cohort design:** All registered patients are followed up for a time period following a (usually random) index date, with their symptom status identified using a lookback period before the index date, and disease risk compared in symptomatic and asymptomatic patients (adjusting for other exposures).

Appropriate controls or comparators should be chosen carefully and the results interpreted with nuance. Simply presenting in primary care for any reason is associated with excess disease risk compared to non-consulting patients of similar age and sex, because of increased disease severity and background morbidity in anyone who presents to primary care; a phenomenon sometimes termed the 'symptom iceberg' (113,114) (Figure 3.2). Comparisons to background disease risk in the general population, adjusting for age and sex, such as that used in a previous study into abdominal pain(48), while providing important context to the observed risk in a symptomatic cohort, can support only limited conclusions about the 'diagnostic value' of the symptom. Instead, comparisons to the background disease risk in patients presenting to primary care (without the symptom)

quantifies the excess risk conferred by specifically presenting with the symptom of interest, as opposed to simply presenting for any reason.

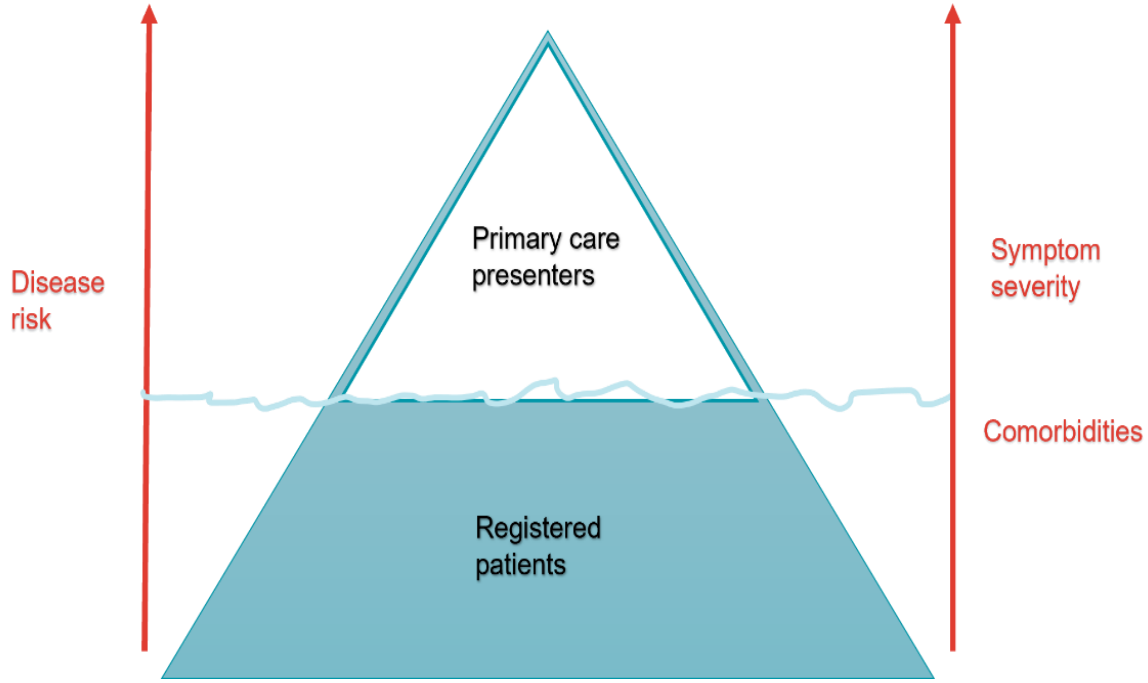



*Figure 3.2. The symptom iceberg in primary care*

Different designs provide varying levels of confidence that a disease symptom is associated with increased disease risk. 'Matched control' and 'reference population' designs that match/adjust for age, sex, and sometimes presentation date and GP practice contextualise risk and establish whether the symptom of interest is associated with increased disease risk. These associations are not necessarily causal, due to the potential for confounding by other factors, such as comorbidity. Adjustment for such confounders can be challenging to implement, as they need to be adequately recorded in the dataset for a satisfactory lookback period. Identifying appropriate confounders can also become complex in studies of multiple disease outcomes, as those that are highly relevant for one disease may not be relevant for another. For example, chronic obstructive pulmonary disease (COPD) may be a highly relevant confounder when comparing risk of lung cancer in patients presenting with or without fatigue, but not when comparing risk of depression. However, for studies aiming to inform clinical practice, it may be sufficient simply to know that a patient's symptom, in combination with their age and sex, indicates that they are at increased disease risk, without knowing the causal process.

### *Comparing between multiple diagnostic outcomes*

Electronic health record based research aiming to help GPs decide which of multiple diseases to investigate in symptomatic patients should consider combining 'descriptive' studies of absolute disease risk with 'diagnostic value' studies of excess risk. A study by Withrow et al is an excellent

example (96). They present hazard ratios for twelve diseases in patients presenting with unexpected weight loss compared to matched controls presenting without, describing the strength of the association between each disease and weight loss presentation. They also present the observed absolute risks of each of the diseases in patients with weight loss, to inform GPs about the order in which to conduct initial investigations.

Descriptive risk studies that rank diseases by their absolute risk in a symptomatic cohort (50,55,98,102) can provide information about the most likely diseases to suspect first. But such a design cannot establish whether the risk of the most likely diseases in the symptomatic cohort is different to what would be expected of these patients. For this, a 'diagnostic value' design is needed, where diseases with the greatest increase in risk in the symptomatic cohort are compared to a comparison group (as in Withrow *et al*(96)).

Figure 3.1 illustrates how using one of these approaches in isolation may provide an incomplete picture of which diseases are most likely in a symptomatic cohort. If absolute risk were considered in isolation, disease 8 would be chosen as the 'most likely' disease (10% absolute risk, 100% relative excess risk), yet the relative excess risk indicates that the symptom has greater predictive value for disease 4 (8% absolute risk; 700% relative excess risk).

### 3.5.3 Choosing appropriate statistical methods to address loss to follow up and competing risks

When using primary care EHRs to study disease risk in symptomatic cohorts, researchers should consider statistical methods that are appropriate for their purposes and mitigate any shortcomings of the data.

The choice of statistical method should consider whether there is loss to follow-up and the reasons why it occurs. In particular, competing risks are events that mean it is impossible for a patient to experience the primary outcome of concern, typically but not necessarily death (e.g., for a study of endometrial cancer risk, hysterectomy would be a competing risk). There may also be issues with informative censoring, where patients are lost-to-follow-up for reasons that relate to their disease risk (e.g., changing practice because they are unhappy with how their symptoms are being handled by their GP), but which do not necessarily lead to or prevent a diagnosis of the disease of interest. Non-informative censoring (115) where patients are lost-to-follow-up for reasons unrelated to their risk of the outcome, may still cause problems either for the sample size or for the accuracy of simple methods of estimating risk.

#### *Loss to follow up*

Patients are sometimes lost to follow up when they change GP practice, which can introduce informative censoring if continued follow up is not available via other datasets (e.g. national registries). The direction of bias is difficult to predict, as patients could change practice for a range of reasons, including dissatisfaction with how the symptoms were addressed, moving to a hospice, or moving home. Mortality rates have been shown to drop below that of the general population after patients have been registered to a Clinical Practice Research Datalink (CPRD) practice for more than two years. This suggests that patients who remain at the same practice throughout a whole follow up period are generally healthier, which could artificially deflate disease risk estimates(116).

#### *Death as a competing risk*

An extremely informative type of censoring is death, as this is an ‘absorbing’ state, which means that patients who die cannot under any circumstances go on to be diagnosed with any disease.

Various relevant studies estimate crude cumulative incidence (48,50,55,96,98,105). In these, risk is calculated as a proportion of patients in the denominator at the start of follow up, regardless of whether they died during follow up. Because patients who died are not censored, death is not a source of informative censoring. This generates risk calculations that are easily interpretable for GPs and reflect a patient’s ‘real’ risk at the time of presentation in a world where they can die of other causes(115). Comparisons between patient groups should be interpreted with caution, however. Disease risk may be lower in some groups of patients because they are at high risk of death (e.g. patients aged 90 years and over), but not because they would be at low risk of the disease if they survived for the entire follow up period. These methods may also be vulnerable to non-informative censoring, if for administrative reasons some groups of patients do not have complete outcome ascertainment.

Some studies censor patients who die during follow up, removing them from the denominator from that time point and so considering them no longer at risk of the disease(96,97). This addresses non-informative censoring, if it is an issue, but does not adequately address competing mortality risk, as it assumes that those who died can be represented by those who did not die. Simply removing patients from the denominator at this point is likely to leave a selectively healthier cohort as follow up progresses. To adequately treat death as a competing risk, other statistical methods are needed such as modelling the subdistribution hazard function(115).

### *Other competing risks*

As well as death, competing risks can also include competing diagnoses, in study designs where follow up ends at the first recorded diagnosis. Such a design may be needed in studies that aim to assess the diagnostic value of a symptom to differentiate between subcategories of a disease. This may require an outcome organised into discrete categories, such as in studies investigating a blood test's ability to differentiate between the most likely primary cancer site or a small number of related diseases (102,105).

It is also possible to explicitly account for competing disease risks by using statistical methods that aim to 'update' the risk of an outcome, after an intermediary event occurs. Multistate models are a powerful tool that can be used to understand the relationship between diagnoses at competing risk(117). However, these complex methods are not well suited to studies that characterise the full disease-signature of a symptom, as the addition of each disease outcome exponentially multiplies the complexity of the analysis.

For studies that aim to describe the risk of many diseases in a symptomatic cohort, or the full 'disease-signature' (96,98), it may be appropriate to ignore competing disease risks, by insisting follow up for a disease of interest continues even if a diagnosis of a different disease is recorded beforehand. For some patients, an initial diagnosis such as anaemia could represent a manifestation or misdiagnosis of a more serious disease (e.g., bowel cancer) diagnosed later, so it would be better to continue follow up to capture the true incidence of all diseases.

### *Parallels with all-cause and net survival*

Above, I discuss issues of estimating disease risk in settings with competing risks such as death. There are obvious parallels with epidemiological studies of cause-specific survival or net survival. These methods aim to make 'more fair' comparisons between groups by removing competing events such as death from the analysis, typically measuring survival in relation to background mortality (118).

Some studies, usually clinical trials, do this in a cause-specific framework, i.e., by examining cancer-specific mortality or cancer-specific survival rather than all-cause survival (e.g. (119–121)). Other studies, particularly epidemiological studies, use a net survival framework (e.g. Ederer II, Pohar Perme methods etc. (118,122) that accounts for background mortality rates (for examples see Arnold et al(123) or Girolamo et al(90)). These net survival methods produce estimates of survival in a hypothetical world where the only cause of death is cancer.

In studies of disease risk in symptomatic patients, a highly similar approach would be to use cause-specific time-to-event methods in symptomatic cohorts which essentially produce risk estimates in a world where patients cannot die. As with studies of survival, the choice of statistical model should be informed by the overall study aims. If the aim is to estimate the actual proportion of patients who will develop a disease in a real-world setting (whether this exceeds a particular referral threshold), net survival approaches are not necessarily appropriate, as they generate disease risk estimates that apply to a hypothetical world where patients cannot die of other causes. Instead, appropriate methods include models that appropriately handle competing risks (e.g. multistate models(124)), Fine-Gray models(125), or measures of crude cumulative risk(115).

### 3.5.4 Defining symptomatic cohorts

#### *Choosing an index date*

Studies of disease risk in symptomatic cohorts that use primary care data usually start analytical follow-up for each patient only after they have been registered to the current practice for at least a year, to discard pre-existing diagnoses added as part of the registration process, which could artificially inflate disease risk estimates (126). Many then select the patient's first occurring symptom as the index date(48,49,96–99), which is intuitively meaningful for clinical audiences. For symptoms that occur frequently in primary care – for example, non-specific symptoms such as fatigue – this could bias the sample by including younger patients who are at lower risk of cancer.

Alternatively, a random symptom could be selected as the index date(50), which should result in an age distribution that is more representative of patients presenting with that symptom in primary care. Selecting a random symptom adds flexibility to ask additional questions, such as how disease risk varies by whether the symptom is new-onset or part of an ongoing series of consultations, or by the amount of time that has elapsed since a previous consultation for that symptom, or by the number of previous symptom presentations.

Landmark approaches could also be used (in 'full cohort' designs), where the same patient can be included multiple times (with appropriate handling of standard errors); or time-varying-exposure approaches, where the patients' entire EHR follow-up is included with a symptomatic status that varies over time. These more complex approaches are rarely used in the existing literature, perhaps in part due to the increased complexity of the statistical methods and difficulties in communicating results to non-academic audiences.

#### *Defining symptom severity*

Ideally, studies of disease risk in symptomatic cohorts should assess what 'severity' is represented by symptom phenotypes based on clinical code lists, and if possible, stratify by symptom severity, as more severe symptoms could be associated with a higher likelihood of more serious disease such as cancer. For some symptoms, this might be possible using clinical codes (for example, the dyspnoea phenotype I use in Chapter 5 includes five Read codes that capture the MRC Breathlessness Scale grades 1-5). However, clinical codes available to GPs to record a single symptom usually include a mixture of such codes as well as broader codes that do not mention severity (the dyspnoea phenotype includes 48 codes in total), so it is unlikely that the severity-specific codes would be complete enough for research. Alternatively, symptom severity could be captured using clinical measurements. For example, Nicholson et al used weight and height recording to quantify the degree of weight loss represented by weight loss Read codes in the Clinical Practice Research Datalink (CPRD)(127). Natural 'pairings' of physical measurements to symptoms are not available for most other symptoms. Fatigue is not objectively measurable, so the potential to study its severity is limited, although study designs that choose a random symptom presentation as an index date could, in theory, use the number of previous presentations of that symptom as a proxy for severity (or at least persistence). Alternative approaches could include whether a symptom presentation was accompanied by other notable healthcare use events such as prescriptions for certain medications.

### *Defining 'disease-free' cohorts*

When describing disease risk in symptomatic patient cohorts, it is typical to exclude patients with a recent diagnosis of that disease in order to minimise the likelihood that the symptom of interest is attributable to a previous diagnosis of the disease of interest, rather than a subsequent diagnosis. It is less clear whether patients with a previous history of the disease should also be excluded, and how to draw distinctions between 'recent' and 'past' diagnoses in patients' EHRs. A long-passed diagnosis is less likely to directly cause a patient's current new-onset symptom, but its existence in the patient's history could be a strong predictor of baseline risk of the disease. For instance, patients with chronic diseases, such as diabetes, HIV, or Inflammatory Bowel Disease (IBD), usually have the disease for life, whereas acute infections (e.g. Urinary Tract Infections (UTIs), influenza) can be diagnosed multiple times in a patient's life, with a prior diagnosis having uncertain influence on a patient's current risk. While diseases such as cancer are not strictly 'chronic', a prior diagnosis can still increase a patient's current cancer risk (128).

Then again, removing patients with previous diagnoses will leave healthier disease-free patients in the cohort, making the study less representative of a real-world population. The longer the lookback period used to identify previous diagnoses (e.g. if patients with any previous history of the disease are removed), the less representative the cohort will be of the 'average' patient who presents to the GP with a symptom (129). In addition, an often-overlooked consequence of using a longer lookback period to define disease-free cohorts is that longer periods of adequate primary care electronic health records (EHRs) are needed for each patient. Patients who have been registered at a practice for more than one or two years have lower mortality risk (116), suggesting that such selection criteria would bias disease risk estimates by selecting a healthier population. The actual decision taken depends on whether the research aims are concerned solely with the detection of new disease. It would be possible for example, to stratify the cohort into patients with no prior disease history, a recent diagnosis, or a past diagnosis.

### 3.5.5 Defining co-occurring features

Patients with non-specific symptoms may also present with co-occurring 'alarm' symptoms, which could alone indicate the presence of underlying disease and explain increased disease risk in cohorts presenting with non-specific symptoms. Studies of non-specific symptom cohorts could better inform clinical practice if they restrict the cohort to patients who did not also present with an alarm symptom, although none have to my knowledge. However, studies can, and have, examined risk in combination with other co-occurring 'non-alarm' symptoms and clinical features, to identify patients at greatest disease risk in a non-specific symptom cohort (49,50,99,110,130)

For both types of studies, there are inherent challenges in identifying symptoms or signs that co-occurred with the index symptom, since in EHRs, patients' continuous states can only be recorded at discrete intervals (when patients present). In studies aiming to describe cancer risk in symptomatic patients or assess the diagnostic value of a symptom (see Section 3.5.1), researchers need a method to decide whether symptoms are co-occurring, usually by choosing an inclusion time window before and after the index presentation to search for records of other symptoms. Current literature inadequately documents the rationale of the exact time window chosen, as well as the potential impact on disease risk estimates (44,45,48–50).

In theory, features recorded after the index symptom should not be used to define cohorts of patients with co-occurring features. This can potentially introduce immortal time bias(131), where patients who 'survive' for longer without being diagnosed with the disease are more likely be included in the co-occurring feature group. In practice, however, a short inclusion period after the index is needed specifically to capture symptoms, because of the possibility of patients not reporting, and doctors not recording all presenting symptoms during the initial consultation.

The time period before the index date should also be as short as possible, as features are more likely to relate to each other (and an underlying disease such as cancer) if they occur close together in time. In contrast, the longer before the index symptom that an accompanying symptom was recorded, the more likely it is to represent a previous diagnostic episode or an unrelated complaint, and the less likely it is to relate to the new subsequent diagnosis of interest.

### 3.5.6 Defining outcomes

#### *Choosing a follow up time period*

Most existing evidence examines disease risk during fixed 'en bloc' periods following a symptom presentation (e.g. within 12 months) (44,45,48–50,55,96–102,105–111). This ignores the fact that disease risk will typically wane over time following an index symptom presentation (96,97). A handful of studies have indeed shown that disease risk in a symptomatic cohort, and the excess risk compared to asymptomatic controls, is sensitive to the follow up time period chosen(96,97).

For instance, if follow-up is very long, *excess* disease risk in symptomatic patients will be under-estimated compared to asymptomatic controls, as background diagnoses will progressively accumulate in cases and controls over time. The *absolute* risk of disease could also be exaggerated in the symptomatic cohort, as a larger number of background cases will have accumulated. Conversely, if the follow-up period is too short, certain diagnoses of which the presenting symptom was truly a prodrome will not be counted, leading to under-estimation of *absolute* risk in the symptomatic cohort. Therefore, studies should supplement crude 'en bloc' incidence estimates with sensitivity analyses showing the impact of the follow up time period chosen. In addition, studies of disease risk using survival analysis methods that rely on a proportional hazards assumption (i.e. that the hazard remains constant throughout the follow up period) should be aware that this does not hold true for symptoms.

#### *Combining data sources to define the outcome*

Full follow up is not always available in primary care for individual patients, if they cannot be tracked between GP practices, or the dataset is limited to GP practices that use a particular computer system. Furthermore, some outcomes more commonly occur outside of primary care so are poorly recorded in primary care EHRs, such as death (132), or diseases that are often diagnosed through referral to secondary care or as emergency hospital admissions, including cancer (133,134), and stroke.

Both issues can be addressed in studies examining disease risk in symptomatic cohorts by using linked datasets. National datasets provide continued follow up, so that censoring is not necessary when patients leave the primary care dataset (135). For instance, in England, cancer registry data alone(136) adequately captures cancer diagnoses, but other diseases that are often recorded only in secondary care but do not have national registries (e.g. stroke) should be identified using both primary and secondary care data. There are questions of how to address conflicts when combining datasets, including whether to prioritise information recorded in datasets with a history of better case ascertainment, or simply prioritise information about the first record of the feature. Best practice is unclear in a setting involving multiple types of data sources and disease outcomes (135).

### 3.5.7 Developing code lists

Features are often recorded in primary care electronic health records (EHRs) using granular and 'messy' coding systems that are poorly suited to epidemiological research(91). In studies of disease risk in symptomatic cohorts, careful phenotyping of symptoms, measurements, test results and diseases is required, though the methods used have often been poorly documented and difficult to replicate (94). Methods to develop phenotypes include:

- Clinician-led: Key terms for the feature of interest are developed through clinical input. Data dictionaries are then searched for the key terms, and reviewed for inclusion. Coding systems with a hierarchical structure, such as Read code v2 or SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms), can also be leveraged to identify relevant sibling or parent codes(94).
- Data-led: Entirely data-led approaches such as cluster analysis (of codes), and natural language processing have been used to generate 'bottom-up' groupings of connected codes into relevant broader concepts. This is more commonly used in EHR systems with access to free text, but examples exist where it has been applied to structured coding systems in the UK (137).
- Translation: Existing code lists from a coding system used in one EHR dataset can be directly translated into another coding system for another dataset, using various approaches including concept mapping, as exemplified by the Observational Health Data Sciences and Informatics (OHDSI) community(138,139).
- Combined approaches: Data-led approaches can be used to improve the sensitivity of clinician-generated code lists ('clinician + data validation'), or conversely, clinicians can improve the specificity of data-generated code lists ('data + clinician validation'). For example, a major resection code lists(140) was generated by identifying the most common procedures occurring shortly after cancer diagnosis, which were then reviewed for specificity by clinicians, who excluded irrelevant procedures.

While large-scale projects are improving access to EHR phenotypes (e.g. CALIBER(141), Cambridge multimorbidity score(142), researchers at Exeter University(143)), phenotypes are often produced for different purposes and in different settings. For example, static code lists generated using data-led approaches can easily become less sensitive over time, as new codes are introduced. Therefore, code list sensitivity and specificity should be checked before being used in a new study (94,95). Another scenario where phenotypes may have limited generalisability or need significant adaptation is in the definition of disease sub-types (e.g. different morphological types of cancer of the same site), as different purposes may require groupings that are more or less coarse.

Finally, some data items are selectively recorded in primary care EHRs. For example, GPs may note a range of symptoms in free text, but only record symptoms of prior concern in coded data, which can bias risk estimates in symptomatic cohorts if the free text is not available to researchers (89). While this example cannot be addressed and must simply be considered at the interpretation stage, in other cases, patients' symptoms may be supplemented with and/ or validated using test results, as shown with unexpected weight loss(97).

## 3.6 Using CPRD to study undetected disease risk in patients with fatigue

### 3.6.1 Defining a study purpose

The overall aims of my primary studies were to generate evidence that can inform diagnostic guidelines and help GPs to appropriately suspect cancer or other possible diagnoses in patients presenting with new-onset fatigue. They can therefore be described as what I term 'descriptive risk' studies (as defined in Section 3.5.1 'Defining a study purpose'), in which I aim to estimate the absolute risk of cancer and other diseases in patients with fatigue, including how this varied by patient demographic characteristics (e.g., age, sex), and other presenting symptoms. However, they also incorporate elements of 'diagnostic value' studies (as defined in Section 3.5.1), as I aimed to contextualise disease risk against background risk in the general population and in patients presenting without fatigue, to assess to what extent fatigue could add discriminatory value in the diagnostic process.

### 3.6.2 Introduction to CPRD

I used CPRD (Clinical Practice Research Datalink) GOLD linked to Cancer Registry (CR), Hospital Episodes Statistics (HES) Admitted Patient Care (APC), and Index of Multiple Deprivation (IMD) datasets. CPRD GOLD stores data about patients registered with participating general practices using Vision® software in the UK. In Box 1, I describe how the strengths and limitations of primary care EHR datasets apply to CPRD GOLD. Several aspects of the methods I have used aim to mitigate these limitations, as explained later in this chapter.

**Box 1. Strengths and limitations of CPRD GOLD for studies of disease risk in symptomatic cohorts**

**Strengths**

- *Coverage. C*overage included 674 GP practices and approximately 6.9% (N= 4.4 million) of the UK population in 2013, providing large sample sizes for cohort studies(145)
- *Representativeness.* Patients were broadly representative of the age, sex, and ethnicity distribution of the UK population in 2013(145)
- *Patient follow up*. Good continuous follow up for most patients; 79 million person-years follow up total, and 9.4 years median follow up per patient, in 2013(145)
- *Overall data quality.* CPRD provide dates between which data is deemed of research quality for patients (acceptability status) and practices (up to standard status), applying multiple validity criteria for both(145)
- *Data richness.* Data available in CPRD include patients' demographic information (age, sex), recorded consultations, symptoms and diagnoses, tests, prescriptions, and referral to other services(145)
- *Data linkage.* Patient-level linkage to other population-level national datasets is possible using NHS number, including cancer registration (CR) data for 'gold standard' ascertainment of cancer diagnoses, Hospital Episodes Statistics (HES) Admitted Patient Care (APC) and Outpatient Care (OP), and Diagnostic Imagine Datasets (DID), and the Index of Multiple Deprivation (IMD)(145)

**Limitations**

- *Recent coverage.* The number of currently participating practices, and therefore current coverage, has fallen over time, from 674 in July 2013(145) to 403 in February 2022(146).
- *Patient follow up between practices.* Patients cannot easily be traced if they change CPRD practices, or move to a practice not included in CPRD.
- *Inaccurate or missing data.* Free text data is not available to researchers in the UK, and there may be selective recording of symptoms and signs in coded data(89). Date of death may be inaccurate for some patient groups (132), and cancer diagnoses may be missing or inaccurate in terms of their date or cancer site (133,134).
- *Phenotyping*. As with many EHRs, phenotypes need to be created for epidemiological research. Centralised resources for disease phenotypes using Read codes exist for CPRD GOLD (141,142).

My empirical studies examining fatigue were part of a group of studies using two data extracts of multi-symptom cohorts generated by CPRD (ISAC protocol 18_299R).

### Extract #1

The empirical studies in Chapters 4 and 5 used a data extract of pre-selected cohorts of patients with a primary care record of at least one of fifteen pre-specified cancer symptoms (including fatigue) between 2007 and 2016 in England, while aged 30-99 years, identified from Clinical Practice Research Datalink (CPRD) GOLD (March 2019 database build). Patients' incident cancers diagnosed from 2006-2015 were extracted from cancer registry data held by the National Cancer Registration & Analysis Service (CR). The coverage of relevant linked data are shown in Figure 3.3. In addition, the patient's neighbourhood Index of Multiple Deprivation (IMD) decile was identified by CPRD, by linking the patient's postcode of residence to its respective Lower Super Output Area (LSOA) in 2011, which was then linked to the LSOA-level 2015 IMD decile. The neighbourhood's IMD decile is derived from a weighted composite score of the area's deprivation across eight domains: housing, employment, income, access to services, education and skills, crime, and living environment(144). Appendix 10.3.1 details the inclusion criteria for the pre-selected cohort, which totalled 1,168,842 patients. For the studies regarding fatigue, a subset of patients presenting with fatigue was identified.

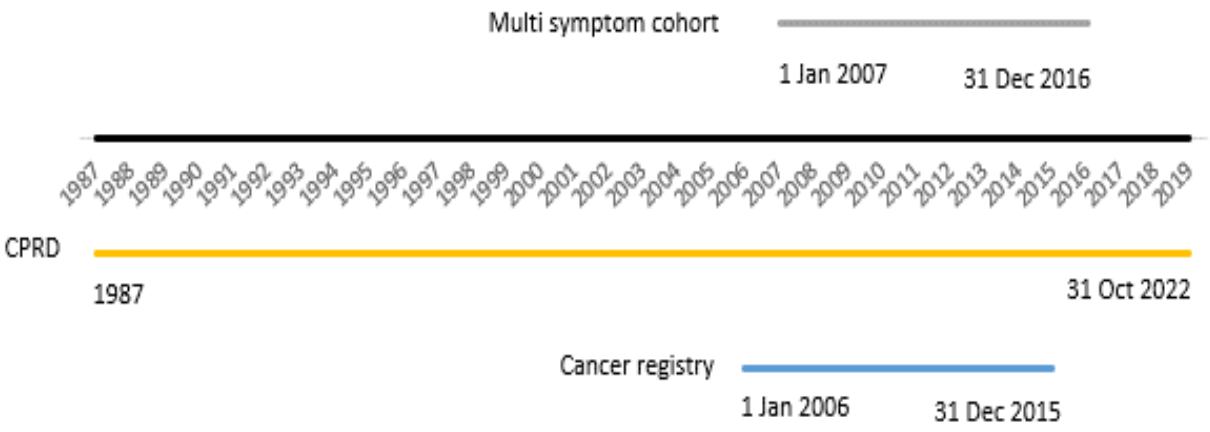



***Figure 3.3. Cohorts of patients identified in CPRD extract #1 and coverage of linked data***

### Extract #2

The empirical study in Chapter 6 used a data extract of pre-selected cohorts of patients with a primary care record of at least one of 22 pre-specified cancer symptoms (including fatigue) and 7 pre-specified test records between 1st January 2007 and 31st October 2021 in England, while aged 30-99 years, identified from Clinical Practice Research Datalink (CPRD) GOLD (November 2021 database build). In addition, a random sample of patients registered to participating practices between 2007 and October 2021 was identified. Patients' incident cancers diagnosed from 1st January 1995 - 31st December 2018 were extracted from cancer registry data, as well as diagnoses

recorded in Hospital Episodes Statistics (HES) Admitted Patient Care (APC) covering 1st April 1997 – 31st October 2020 (Figure 3.4). Appendix 10.3.2 details the inclusion criteria for the pre-selected cohorts, which totalled 2,530,253 patients for the symptomatic cohorts, and one million for the random sample, of which subsets were included in the study on fatigue.

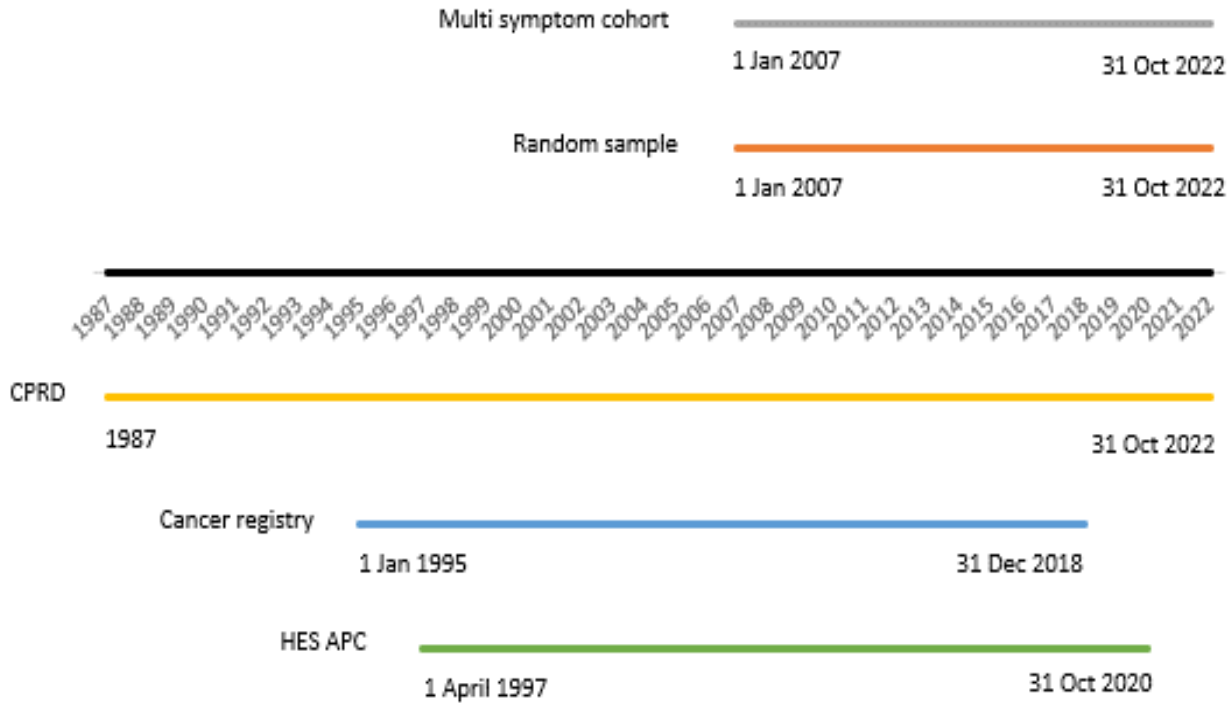



*Figure 3.4. Cohorts of patients identified in CPRD extract #2 and coverage of linked data*

For both data extracts, HES and CR data were identified by CPRD using a pre-existing eight-step deterministic linkage algorithm including NHS number, sex, date of birth, and postcode. The Index of Multiple Deprivation (IMD) decile of the patient's neighbourhood of residence was also identified, through linkage via the patient's postcode.

## Data storage and management

The two CPRD data extracts were received via a secure file transfer portal (SFTP) and stored within UCL's Data Safe Haven (DSH). The Data Safe Haven has been certified to the ISO27001 information security standard and conforms to NHS Digital's Information Governance Toolkit.

The full dataset received from CPRD for extract #1 totalled approximately 125 GB of data spread over 354 data files. Extract #2 totalled approximately 251 GB of data spread over 406 data files. I conducted a preliminary data management step using PuTTY version 0.73, in which I appended data tables that were received as multiple files.

The majority of data management was conducted in MySQL Workbench version 6.1. Data were stored as SQL data files (.dta) accessed via MySQL Workbench, with accompanying SQL code files

(.sql) recording how they were loaded into the database. For each extract, 26 data files (Figure 3.5), and their accompanying lookup files were loaded into SQL.

Additional data management, and all statistical analysis was conducted in Stata versions 16 and 17, and R version 4.1.2, once relevant tables had been selected. Tables were read directly into Stata and R using MySQL Open Database Connectivity (ODBC) 8.0 Unicode Driver.

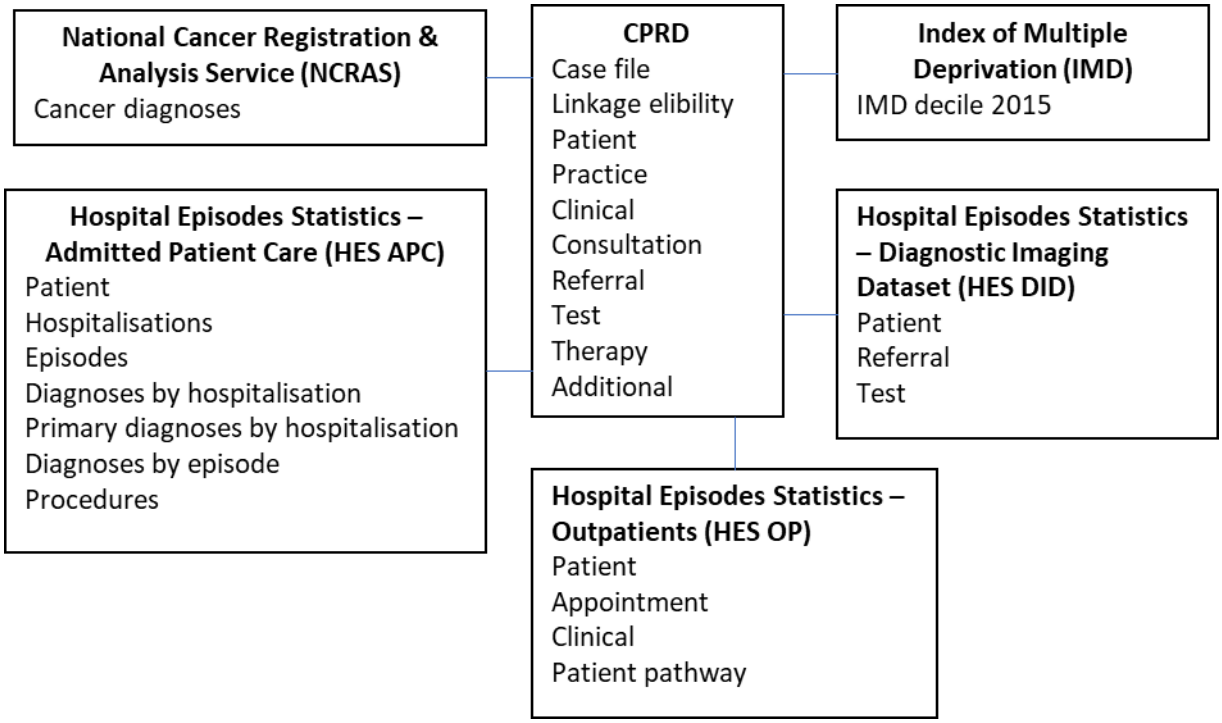



*Figure 3.5. Structure of data included in extracts #1 and #2*

## Ethics approval and consent to participate

This study was approved by the UK Medicines and Healthcare products Regulatory Agency (MHRA) Independent Scientific Advisory Committee (ISAC Protocol number 18_299RMnA5), under Section 251 (NHS Social Care Act 2006). This study is based on data from the CPRD obtained under licence from the MHRA. The data is provided by patients and collected by the NHS as part of their care and support.

### 3.6.3 Choosing appropriate comparisons

*Comparing against baseline risk*

I chose to compare disease risk in fatigue presenters to the background risk in the general population (or registered patients), similarly to a previous study into abdominal pain(48), and also to primary care patients presenting without fatigue, similarly to previous studies into weight loss (49,96,97). Using these comparisons, I aimed to situate fatigue within its symptom iceberg. I hypothesised that if disease risk in fatigue presenters was similar to non-fatigue presenters, but both were higher than in registered patients, then excess risk in fatigue presenters could be associated with simply presenting in primary care.

I did not aim to fully explain the association, but to use comparisons to contextualise disease risk in a symptomatic cohort. Therefore, I adjusted for and/ or stratified by age and sex, but did not adjust for other confounders, such as lifestyle factors and comorbidities. Hence associations between fatigue and increased risk of underlying disease are not necessarily causal (i.e., biological disease processes directly link fatigue with the underlying pathology). For instance, if we observed excess bowel cancer risk in fatigue presenters, we would not assume that fatigue is causing bowel cancer, but that bowel cancer is causing fatigue. This could be via a direct pathway, whereby undetected bowel cancer directly causes patients to feel fatigued. It could also be via an indirect pathway, for example, undetected bowel cancer could trigger anaemia, which itself causes fatigue (Figure 3.6). This mediator is only problematic if comparing the relative risk of different diagnostic outcomes in fatigued patients, where bowel cancer and anaemia are two possible outcomes.

It is also plausible that some of the observed association is due to confounding. For example, obesity could cause both fatigue and bowel cancer, and perhaps some of the observed association could be due to this common cause. If this was the case, then the presence of fatigue might indicate that patients have a higher baseline risk of bowel cancer, but the symptom might not necessarily be a strong signal of present but currently undetected cancer.

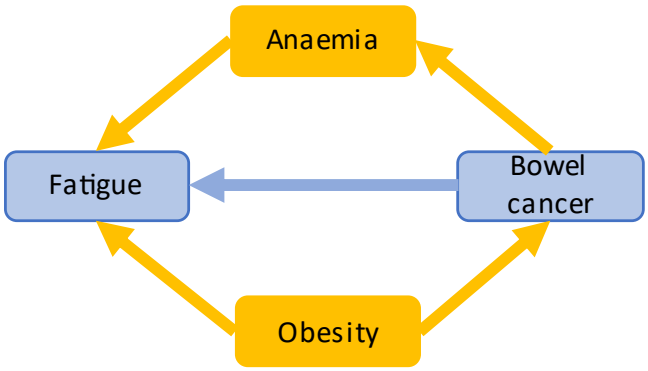



***Figure 3.6. Hypothetical diagram illustrating a potential mediator on the causal pathway between cancer and fatigue (anaemia) and a potential confounder (obesity)***

### *Comparing between multiple diagnostic outcomes*

To inform GPs about the most likely diagnoses to suspect first, I described the *absolute* risk of different diseases (including specific cancer sites) in patients with fatigue (a 'descriptive risk' design as defined in Section 3.5.1). In Chapter 6, I also incorporated elements of 'diagnostic value' studies, by identifying diseases with the greatest *excess* risk in fatigue presenters compared to non-fatigue presenters or the general population, and ranking diseases by this excess risk. These comparisons were most informative for diagnostic guidelines, as they identified which diseases had the strongest associations with fatigue, rather than simply reflecting which were most common in general in primary care.

### 3.6.4 Choosing appropriate statistical methods

#### *Loss to follow up*

In my studies, although patients could only enter while they were registered to an up to standard Clinical Practice Research Datalink (CPRD) practice, some patients were then lost to follow up (LTFU) in CPRD due to changing GP practice.

One option to address this was to remove LTFU patients from the study altogether if they left CPRD during the one year follow up period. I discounted this option because changing GP practice is not a random event (116); it could be a proxy for poor health or death (soon after disenrollment with a specific practice), so removing these patients could leave an unrepresentative, healthier, cohort. Another option was to include LTFU patients in the initial study cohort, but censor them from the at-risk denominator at the point that they became LTFU in CPRD, using time-to-event methods as per some previous studies (96,97). This option was also unsuitable for the same reason (i.e. because as changing GP practice is not a random occurrence, and selects a progressively healthier cohort as follow up continues).

Instead, I chose to calculate crude cumulative incidence, where risk was calculated as a proportion of patients in the denominator at the start of follow up, regardless of whether they became LTFU. This approach ensured that the study cohort represented all patients presenting with fatigue in primary care. It was suitable as full case ascertainment of cancers was possible using linked national cancer registry data, and the loss of ascertainment of other diagnoses because of LFTU is substantially mitigated by the inclusion of secondary care (HES APC) data. However, the risk of diseases that are predominantly diagnosed in primary care (e.g. depression) could still be underestimated in patients who were LTFU in CPRD, compared to cancers or diseases that are commonly recorded in secondary care. Such underestimates could be exacerbated in groups of patients who frequently change GP practice.

#### *Death as a competing risk*

I estimated crude cumulative incidence, that is, risk as a proportion of patients in the denominator at the start of follow up, regardless of whether they died during follow up. This was because follow up was relatively short (up to 1 year), so the impact of death as a competing event would be minimal. It also supported risk calculations that were easily interpretable for GPs and reflected a patient's 'real' risk shortly after an initial consultation(115). This means that although disease risks estimated in my thesis are more relevant to clinical practice, they should be interpreted with caution in patients at high risk of death (e.g., aged 90 years and over). In these groups, disease risk may be low because death commonly occurs first, rather than because they would be at low risk if they survived for the full follow up period.

#### *Other competing risks*

As I aimed to describe the risk of many diseases (fatigue's 'disease-signature'), I ignored competing disease risks, and continued follow up for a disease of interest even if a diagnosis of a different disease was recorded beforehand. This best reflected the real-world risk of each disease, including those that are often diagnosed after a delay or initial misdiagnosis.

### 3.6.5 Defining a symptomatic cohort

#### *Choosing an index date*

I aimed to identify a cohort of patients presenting to primary care with new-onset fatigue, to inform GPs about what steps to take at a patient's initial consultation, and to minimise the likelihood that the patient's fatigue was attributable to a previously diagnosed condition or disease (including cancer). Therefore, I identified the patient's first 'eligible' record of fatigue, that is, the first record that met other study criteria such as being when the patient was age 30 years. This is illustrated in Figure 3.7; patients A and B both had fatigue records after the point that they entered the study (entry was at year 0), and before the point that they left the study (which was at year 3). The first of their 'eligible' fatigue records was chosen in both cases.

I also excluded a small group of patients who had an 'ineligible' record of fatigue one year before their first eligible fatigue record, for example, because it was recorded before the patient was 30 years old (patient D in Figure 3.7). However, if such a patient had another eligible fatigue record over a year later, the later record was selected and the patient was included (patient C in Figure 3.7). This meant that patients did not enter the study midway through a series of consultations for fatigue, ensuring the cohort contained only patients with 'new-onset' fatigue.

#### *Defining symptom severity*

It would be helpful to know what 'severity' is represented by the fatigue phenotype captured by clinical codes, and if possible, stratify by severity, as more extreme fatigue could be associated with a higher likelihood of more serious disease such as cancer. This was not possible, as severity was not captured by the code list, and there are no clinical severity scores in use for fatigue. In addition, as I began follow up with the first presentation with new-onset fatigue, I could not use the number of previous fatigue presentations as a proxy for severity (or at least persistence).

#### *Defining 'disease-free' cohorts*

To minimise the likelihood that fatigue was attributable to a previous diagnosis of cancer, for my first primary study, I also excluded patients if there was a cancer diagnosis recorded in the cancer registry in the year before or on the same day as their first eligible fatigue record (patient E in Figure 3.7). In these cases, the patient was still included if they had another eligible fatigue record more than a year after the first eligible fatigue record. A sensitivity analysis examined the impact on cancer risk of extending the look-back period from 1 to 2 years, and also where these two exclusions were not conducted.

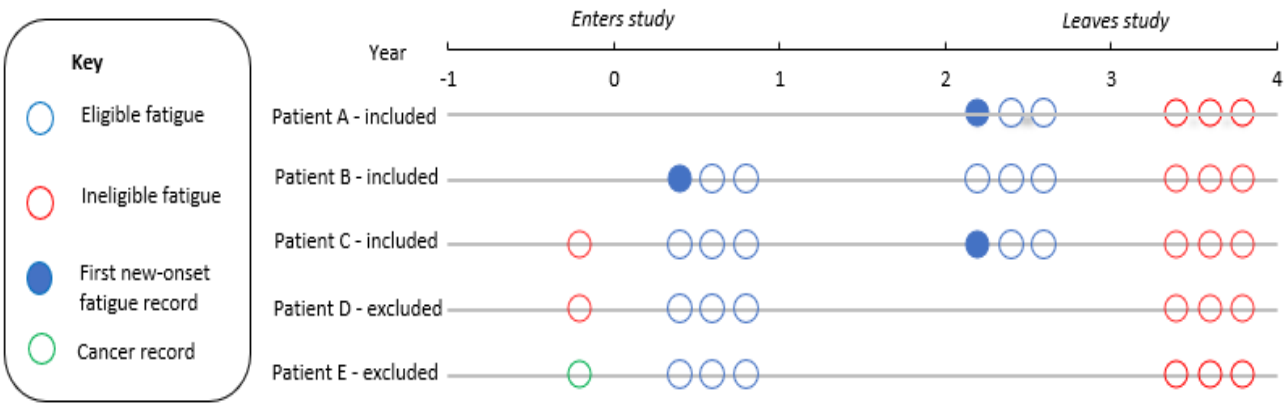



*Figure 3.7. Identification of new-onset fatigue - illustrative scenarios*

If the patient had a previous diagnosis of cancer less recently (e.g. 2-15 years ago), it is less likely to directly cause the new-onset fatigue, but its existence in the patient's history could be a strong predictor of subsequent disease diagnosis. In particular, previous diagnoses of chronic diseases (e.g. Inflammatory Bowel Disease (IBD), HIV etc.) will be highly predictive of subsequent diagnosis. Therefore, in the third primary study, I also excluded patients from the cohort for each disease if they ever had a previous diagnosis of that disease. I made exceptions for some infections that are likely to occur multiple times in a patient's life; for these, patients were only excluded from the cohort for that disease if they had a previous diagnosis of it in the previous two years before their first fatigue record. I conducted a sensitivity analysis to identify the impact on disease risk of including all patients, regardless of whether they had a previous diagnosis of each disease.

Due to this inclusion criteria, only patients who were registered to their practice for at least one year (in Chapters 4 and 5) or two years (in Chapter 6) before their index date could be included. As this may have introduced bias into the sample by selecting a healthier population at lower mortality risk(116), in Chapter 6, I conducted a sensitivity analysis to assess its potential impact on disease risk estimates.

### 3.6.6 Defining co-occurring features

It was possible that alarm symptoms for cancer (e.g. breast lump, blood in stool etc.) often accompanied fatigue, and that excess disease risk in fatigue presenters could be associated with these symptoms instead. In Chapter 5, I restricted the cohort to patients presenting with fatigue in the absence of alarm symptoms for cancer, and compared disease risk for patients with and without other co-occurring 'non-alarm' symptoms, to further elucidate risk in this particularly difficult to diagnose cohort. The accuracy of these estimates depended on maximising the sensitivity of the code lists used to define 'alarm' symptoms, to reduce the possibility that patients with alarm symptoms were included in the cohort. Therefore, I meticulously identified a long list of alarm symptoms using NICE urgent referral Guidelines for suspected cancer(16), and supplemented published code lists for these symptoms using both clinician-led and data-led approaches.

When identifying symptoms or signs that co-occurred at the same time as the first fatigue presentation, I chose to include co-occurring symptoms if they occurred three months before to one month after the first fatigue presentation. I included symptoms occurring up to one month after first fatigue presentation (censoring on cancer diagnosis), because of the possibility of doctors not recording all presenting symptoms during each consultation, but did not extend longer than one month to avoid introducing immortal time bias. Figure 3.8 illustrates different scenarios in which a patient's other symptoms would be defined as co-occurring or non-co-occurring with fatigue.

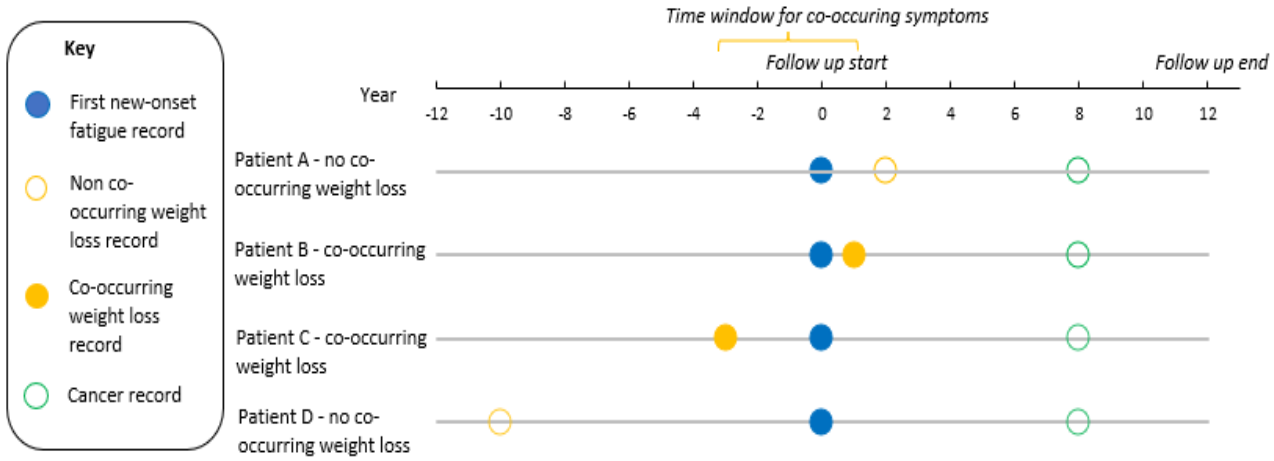



*Figure 3.8. Illustrative scenarios of a co-occurring and non-co-occurring symptom*

Exploratory analysis indicated that including a long time window before the fatigue index date for capturing co-occurring symptoms would generate larger cohorts, and therefore more precise risk estimates (Figure 3.9). In contrast, using a very short inclusion window before the index date would capture fewer patients with fatigue and a co-occurring symptom; fewer than 700 males had co-occurring weight loss 1 month before to 1 month after their first fatigue presentation. This could generate imprecise risk estimates.

However, symptoms recorded a long time before the index date are clearly less likely to be related to the index fatigue presentation and any underlying cancer (if present). This is shown by the lower cancer risk observed in cohorts of patients with fatigue and co-occurring weight loss when a longer time window before the first fatigue presentation was used (e.g. 12 months before to one month

after the first fatigue presentation) compared to a shorter period (e.g. 1 month before to 1 month after) (Figure 3.10). Therefore, including three months pre-index date delivered a good balance between these considerations, with clinical colleagues (MR, YL, CR) confirming that in practice, GPs would tend to look back this far in a patient's history.

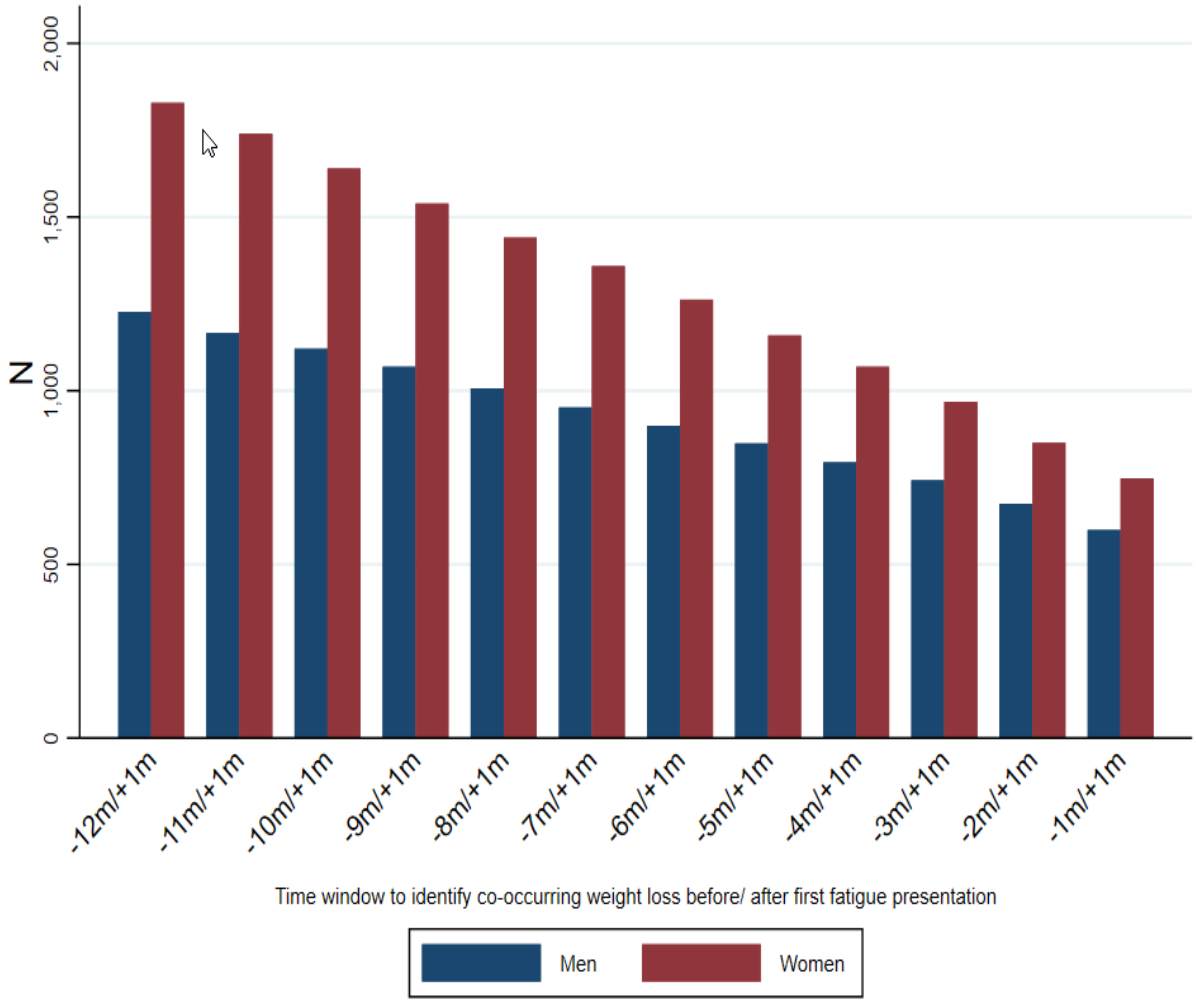



***Figure 3.9. Cohort size (N) of patients with fatigue and co-occurring weight loss, according to the lookback period used to identify co-occurring weight loss before the patient's first fatigue presentation, by gender***

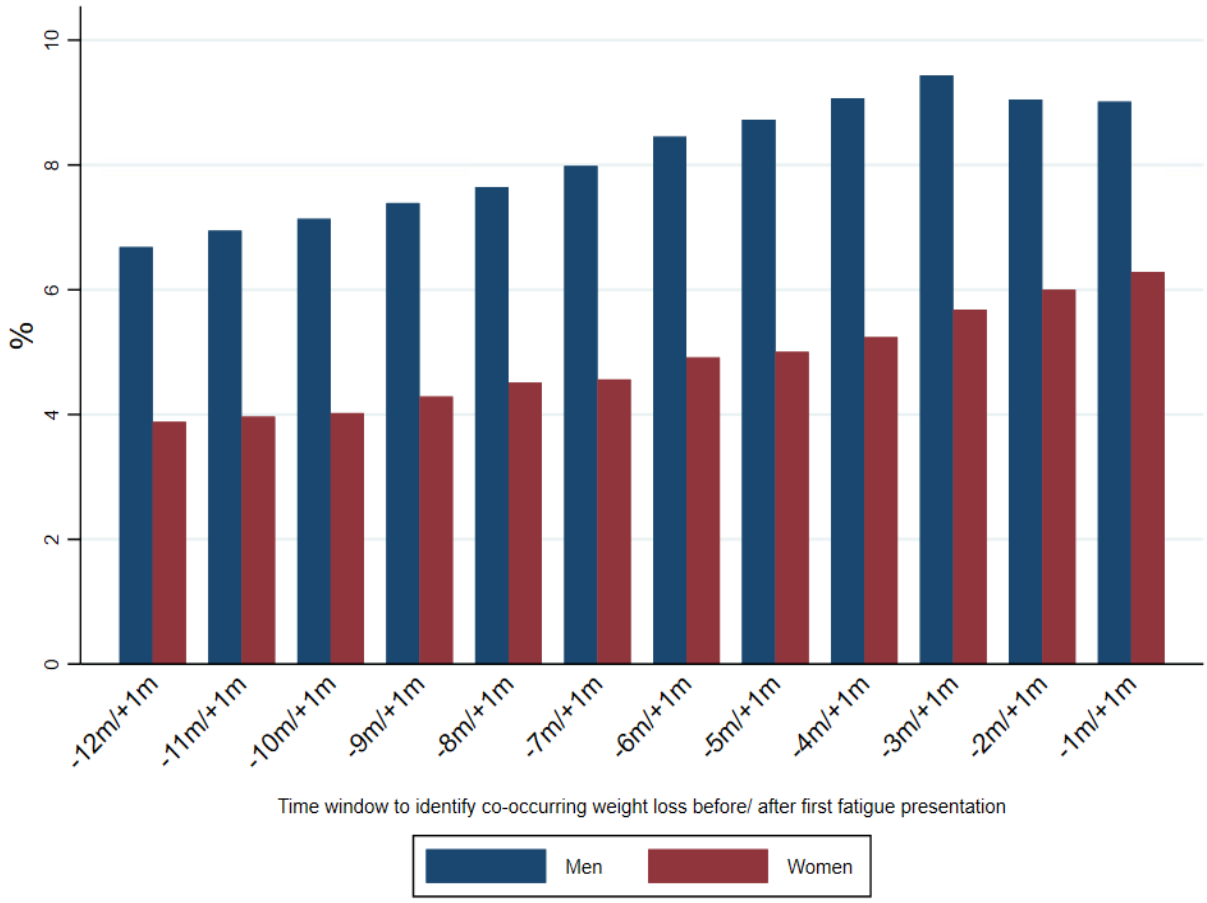



***Figure** 3.10. **Nine-month cancer risk (%) for patients with fatigue and co-occurring weight loss, according to the lookback period used to identify co-occurring weight loss before the patient's first fatigue presentation, by gender***

### 3.6.7 Defining outcomes

*Choosing a follow up time period*

To generate findings that were easily interpretable for clinicians and guideline policy makers, I calculated 'en bloc' estimates of disease risk within a specified period of time (e.g. 12 months in the first and third study) following the index date. However, as disease risk is sensitive to the time period chosen(96), in the third study, I provided supplementary analyses of monthly cumulative risk following the index date. This showed that excess risk was concentrated in a fairly short period (within 3-6 months) for some diseases (e.g. lung cancer, hypo/hyperthyroidism). For these diseases, using a longer 12-month follow up period likely underestimated the excess risk in fatigue presenters relative to their background disease risk at 12-months. In general, the longer the follow-up among patients exposed to a symptom, a) the higher the absolute risk of disease will be, as background cases unrelated to the presenting symptom will accumulate; but also b) the greater the risk of underestimating excess risk compared to controls will be, as background cases will also accumulate in controls. In the second study, only a 9 month follow up period was used, as it focussed only on cancer risk, and the first study indicated that the period of excess cancer risk was concentrated in this period.

*Combining datasets to define the outcome*

Due to the availability of continued follow up at a population level via Hospital Episodes Statistics Admitted Patient Care (HES APC) and National Cancer Registration and Analysis Service (CR) data, I continued follow up for a patient after they left the Clinical Practice Research Datalink (CPRD). In the third study, which I combined diagnoses recorded in CPRD, HES APC, and CR data by taking the date of the first recorded diagnosis of each disease. Cancer diagnoses were an exception; these were taken only from CR, as it is considered the 'gold standard' for identifying confirmed diagnoses and the definitive primary cancer site(145).

### 3.6.8 Developing code lists

*Developing symptom and outcome code lists*

Throughout the empirical studies, I identified other symptoms in primary care that co-occurred with fatigue, as well as a range of disease outcomes including cancer. Similarly to other electronic health record (HER) databases, features are recorded in Clinical Practice Research Datalink (CPRD) and linked Hospital Episodes Statistics (HES) and National Cancer Registration and Analysis Service (CR) data using granular coding systems (Read Codes V2, International Classification of Diseases 10th Revision (ICD 10), and Office of Population Censuses and Surveys Classification of Interventions and Procedures version 4 (OPCS4)) that require grouping into broader concepts for research. Many symptoms and diseases are already well phenotyped in these datasets, due to prior large scale phenotyping projects (141–143). Other published phenotypes exist, but are not centralised, therefore, I gathered existing phenotypes from a range of known sources, and based the features used in my studies on these.

I assumed the previously published code lists had adequate specificity, as they were developed using clinical input, and used in prior peer-reviewed publications. However, code lists can easily become less sensitive over time, as new codes are introduced. I checked the sensitivity of code lists using a previously documented data and clinically-led approach(140), whereby the most common symptom or disease codes in the cohort of interest were identified and then reviewed by clinicians. Where available, I also merged multiple code lists for the same phenotype.

*Quality assuring the fatigue code list*

As my empirical studies were part of a group of studies using a pre-selected cohort of patients with at least one of various pre-specified cancer symptoms (including fatigue), the cohort of patients with fatigue had previously been defined for extract #1 using a list of Read codes collated using methods developed by Hamilton and Price(143).

As it would form the basis of the main study cohort, I quality assured the fatigue phenotype, checking for example that all possible codes for fatigue that featured in the current dataset were captured in the code list, and that coding practices had not changed significantly throughout the study period. I conducted the quality assurance exercise using CPRD extract #1, which captures patients with a record of fatigue between 2007-2016, and improvements to the fatigue code list were applied in the data specification for extract #2.

The original list of Read codes used to define fatigue in extract #1 are available in Appendix 10.3.3. These include diagnostic codes for chronic fatigue syndrome (CFS) and post-viral fatigue (PVF). I retained these in the list for my primary studies, as the aim was to identify the patient's first fatigue-related presentation to define a cohort of patients presenting with new-onset fatigue. Therefore, if a patient's first fatigue-related code was for CFS or PVF, this was likely to be the patient's first presentation with fatigue, but the GP could have made an error when choosing the specific fatigue-related code. However, in the third primary study, different diagnostic outcomes were examined, including "Postviral fatigue syndrome, neurasthenia and fibromyalgia" – a code concept that includes CFS and PVF. In this study, patients with any of the CFS or PVF codes in the fatigue code list

on their index date were removed from the fatigue-presenter cohort when examining the incidence of “postviral fatigue syndrome, neurasthenia and fibromyalgia”.

### Sensitivity of the fatigue code list

Firstly I aimed to ascertain whether the Read codes used to select this cohort captured all relevant coded fatigue presentations in the Clinical Practice Research Datalink (CPRD), using the steps detailed in Figure 3.11.

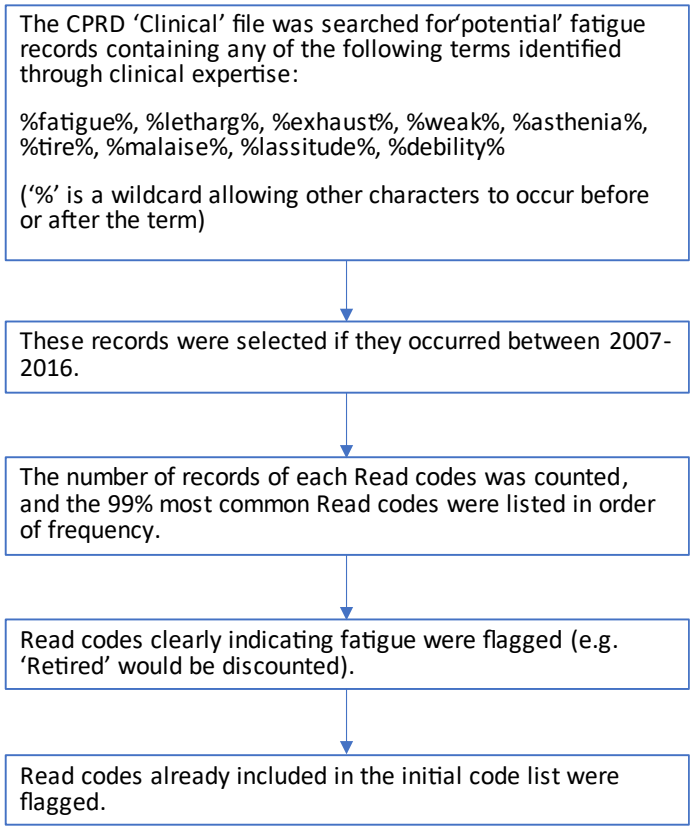



***Figure 3.11. Steps to search for additional fatigue Read codes***

There were 106 ‘potential’ fatigue Read codes present in the dataset for the pre-selected cohort of patients presenting with one of fifteen symptoms, totalling 561,618 fatigue records. 38 Read codes accounted for the 99% most common codes (n = 558,694 records), of which 28 (n = 534,088 records) were classed as ‘fatigue’ on review. Only one of these 28 Read codes was not already included in the initial code list used to define fatigue (168..13 ‘Malaise – symptom’), which accounted for 3.6% (n = 20,108) of potential fatigue records (Table 3.1).

***Table 3.1. Read codes included in definition of fatigue***

*Records of the most common* potential fatigue Read codes as a proportion of all potential fatigue records**, and classification of fatigue Read codes after review*

| Read code | Read code description | Number (n) of records | Proportion (%) of potential fatigue records | Cumulative proportion (%) of potential fatigue records | Included in initial code list | Classed as fatigue after review |
|---|---|---|---|---|---|---|
| **168..00** | Tiredness symptom | 138,895 | 24.7 | 24.7 | Yes | Yes |
| **1683** | Tired all the time | 119,009 | 21.2 | 45.9 | Yes | Yes |
| **R007500** | [D]Tiredness | 38,874 | 6.9 | 52.8 | Yes | Yes |
| **1683.11** | C/O - 'tired all the time' | 32,284 | 5.7 | 58.6 | Yes | Yes |
| **168..12** | Lethargy - symptom | 31,101 | 5.5 | 64.1 | Yes | Yes |
| **1682** | Fatigue | 26,850 | 4.8 | 68.9 | Yes | Yes |
| **1B3..12** | Weakness symptoms | 25,834 | 4.6 | 73.5 | Yes | Yes |
| **E205.12** | Tired all the time | 24,440 | 4.4 | 77.9 | Yes | Yes |
| **168..11** | Fatigue - symptom | 22,905 | 4.1 | 81.9 | Yes | Yes |
| **168..13** | Malaise - symptom | 20,108 | 3.6 | 85.5 | No | Yes |
| **R007000** | [D]Malaise | 15,652 | 2.8 | 88.3 | Yes | Yes |
| **R007300** | [D]Lethargy | 13,031 | 2.3 | 90.6 | Yes | Yes |
| **13J5.00** | Retired | 8,720 | 1.6 | 92.2 | No | No |
| **1681** | Not tired | 4,635 | 0.8 | 93.0 | No | No |
| **F286.11** | CFS - Chronic fatigue syndrome | 4,217 | 0.8 | 93.8 | Yes | Yes |
| **1B32100** | Weakness of leg | 3,513 | 0.6 | 94.4 | No | No |
| **R007100** | [D]Fatigue | 2,878 | 0.5 | 94.9 | Yes | Yes |
| **F286.00** | Chronic fatigue syndrome | 2,764 | 0.5 | 95.4 | Yes | Yes |
| **1684** | Malaise/lethargy | 2,260 | 0.4 | 95.8 | Yes | Yes |
| **1688** | Exhaustion | 2,041 | 0.4 | 96.2 | Yes | Yes |
| **R2y3.00** | [D]Debility, unspecified | 2,017 | 0.4 | 96.5 | Yes | Yes |
| **1B32300** | Facial weakness | 1,631 | 0.3 | 96.8 | No | No |
| **1B32.00** | Weakness present | 1,555 | 0.3 | 97.1 | Yes | Yes |
| **F380.00** | Myasthenia gravis | 1,501 | 0.3 | 97.3 | No | No |
| **R007z11** | [D]Lassitude | 1,480 | 0.3 | 97.6 | Yes | Yes |
| **1B32000** | Weakness of arm | 1,425 | 0.3 | 97.9 | No | No |
| **13JH.00** | 'Retired' - investment income | 1,234 | 0.2 | 98.1 | No | No |
| **1684.11** | C/O - debility - malaise | 1,201 | 0.2 | 98.3 | Yes | Yes |
| **168Z.00** | Tiredness symptom NOS | 1,180 | 0.2 | 98.5 | Yes | Yes |

| | | | | | | |
|---|---|---|---|---|---|---|
| **F286.12** | Postviral fatigue syndrome | 902 | 0.2 | 98.7 | Yes | Yes |
| **F222.11** | Left sided weakness | 824 | 0.1 | 98.8 | No | No |
| **R007211** | [D]General weakness | 715 | 0.1 | 98.9 | Yes | Yes |
| **E205.11** | Nervous exhaustion | 680 | 0.1 | 99.1 | Yes | Yes |
| **F223.11** | Right sided weakness | 585 | 0.1 | 99.2 | No | No |
| **13P..00** | Retirement pensions | 538 | 0.1 | 99.3 | No | No |
| **R007200** | [D]Asthenia NOS | 522 | 0.1 | 99.4 | Yes | Yes |
| **E205.00** | Neurasthenia - nervous debility | 357 | 0.1 | 99.4 | Yes | Yes |
| **2832.12** | O/E - weakness | 336 | 0.1 | 99.5 | Yes | Yes |
| **Total potential records included in the top 99% most common Read codes** | | | 558,694 | | | |
| **Total potential fatigue records 2007-2016*** | | | 561,618 | | | |
| ***Only the top 99% most common Read codes are shown **Data for pre-selected cohort of patients presenting with one of fifteen symptoms in CPRD, between 2007-2016.** | | | | | | |

## Stability of the fatigue code list over time

My examination of the frequency of the Read codes used to define fatigue over time indicated that the recording of fatigue was relatively stable over the study period, and suitable for use in my studies.

I examined the frequency of patients with a 'potentially eligible' fatigue record by year. 'Potentially eligible' records were those meeting criteria detailed in Step 2 of the inclusion criteria in Chapter 4 (see section 4.6.2), with a Read code included in the initial list used to define fatigue. Initial inspection showed the numbers of patients in the overall pre-selected cohort of fifteen symptoms decreased year on year. This is likely because the number of currently practices participating in Vision software (and therefore captured in CPRD Gold), has fallen over time, from 674 in July 2013(145) to 403 in February 2022(146). Therefore, the number of patients with an 'eligible' record of fatigue each year was calculated as a proportion of patients who were eligible for inclusion for all or part of the year. As a proportion of patients in CPRD Gold each year, the number of patients with at least one 'potentially eligible' fatigue record each year decreased slightly, from 5.4% in 2007 to 4.2% in 2016 (Table 3.2).

***Table 3.2. Number and proportion of patients with a 'potentially eligible'* fatigue record, by study year***

| Study year | Fatigue records | Patients with fatigue | | Patients in CPRD** |
|---|---|---|---|---|
| | n | n | % | N |
| 2007 | 56,263 | 47,718 | 5.4 | 885,109 |
| 2008 | 57,711 | 48,856 | 5.3 | 922,027 |
| 2009 | 58,732 | 49,965 | 5.3 | 941,014 |
| 2010 | 55,054 | 46,716 | 4.9 | 947,574 |
| 2011 | 54,881 | 46,843 | 5.0 | 929,381 |

| | | | | |
|---|---|---|---|---|
| 2012 | 50,365 | 42,980 | 4.8 | 892,974 |
| 2013 | 46,308 | 39,652 | 4.7 | 849,299 |
| 2014 | 40,257 | 34,668 | 4.6 | 755,986 |
| 2015 | 31,488 | 27,203 | 4.4 | 618,781 |
| 2016 | 20,802 | 18,240 | 4.2 | 438,011 |

*Potentially eligible records were those occurring between 2007-2016 that met criteria detailed in Step 2 of Section 4.6.2

** Patient had at least one day 'up to standard' follow up at a CPRD practice in that year

Finally, I examined the relative frequency of Read codes by year. Due to the declining numbers of GP practices participating in CPRD Gold over time (and therefore the number of eligible patients), the number of patients with a 'potentially eligible' record of each fatigue Read code per year was calculated as a proportion of patients who had a 'potentially eligible' record of fatigue that year.

The most common Read codes between 2007 and 2016 were 168.00 'Tired all the time' (138,895 records) and 1683 'Tiredness symptom' (119,009 records) (Table 3.1).As a proportion of patients who had a 'potentially eligible' fatigue record each year, the relative frequency of these two codes increased slightly, from 22.6% in 2007 to 27.0% in 2016 for 168.00, and 26.5% in 2007 to 30.2% in 2016 for 1683. In contrast, there was a decrease over time in the proportion of patients with other rarer Read codes. For example, 8.8% of patients with a 'potentially eligible' record of fatigue in 2007 had at least one record of R007500 '[D] Tiredness', but this decreased to 5.8% by 2016 (Figure 3.12).

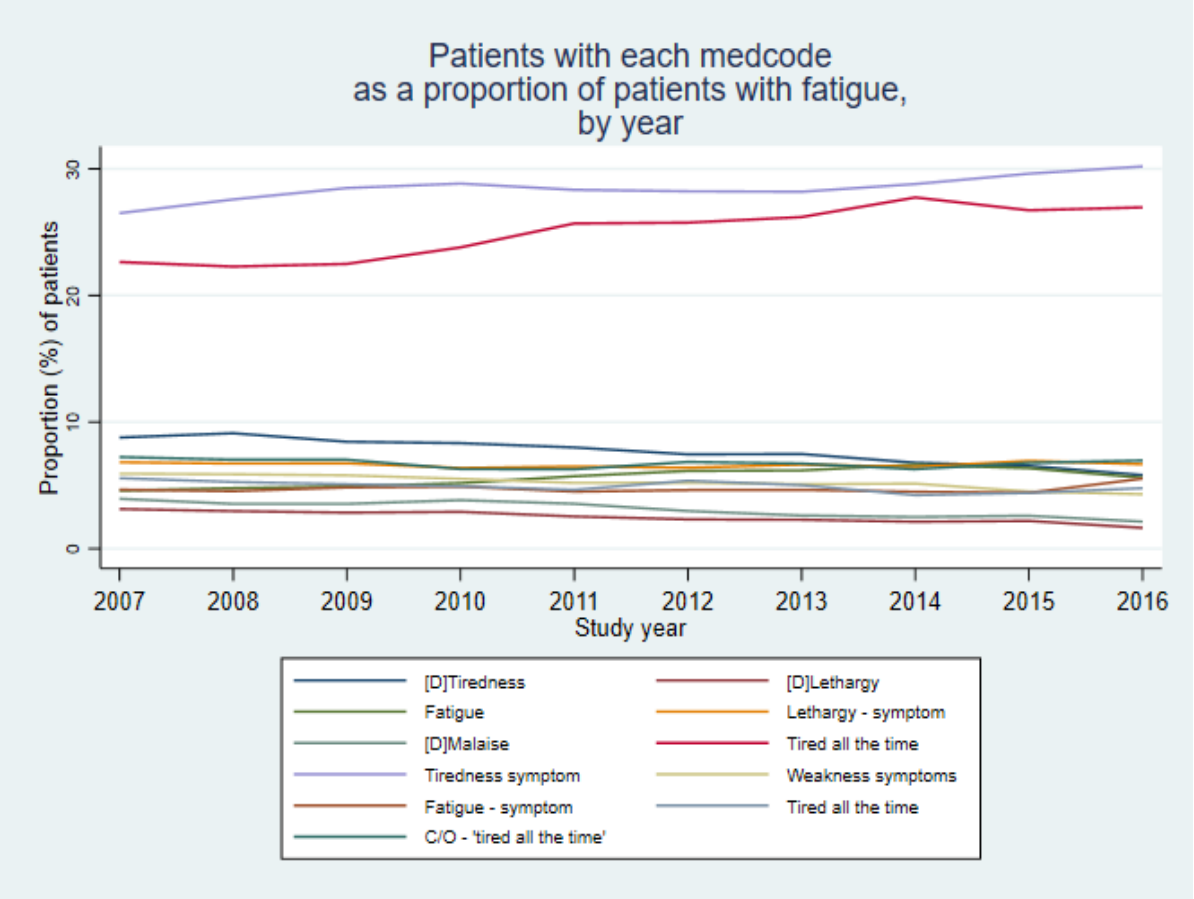



*Figure 3.12. Patients with each Read code, as a proportion of patients with fatigue, by year*

### Overall suitability of the fatigue code list

The analysis showed that the majority of key Read codes needed to select patients with fatigue were captured in the initial code list definition used in the CPRD extract, with the exception of 'Malaise – symptom' (which account for a small proportion (3.6%) of potential fatigue records). Although this analysis did not use the full CPRD dataset, the data extract included full health records for approximately 1.2 patients with 14 other symptoms, so it can be assumed that this would highlight any other commonly used fatigue Read codes not included in my definition of fatigue. As it was not used to define the initial cohort for extract #1, and it accounted for a small number of patients, the extra Read code found ('Malaise – symptom') was not added to the code list definition of fatigue for studies using extract #1. However, it was subsequently added to the code list defining patients with fatigue in extract #2.

The analysis of the pre-selected cohort confirms that the number of patients at participating practices in CPRD Gold decreased over time, and it is unlikely that this would be particular to patients presenting with fatigue (and the other fourteen symptoms) in this cohort. Taking this trend into account, the actual frequency of coding of fatigue for patients participating in CPRD appeared to be relatively stable over time, although there was a gradual decrease. The use of particular Read codes themselves appeared to be relatively stable, with an apparent slight consolidation over time into increased usage of the two most common codes. Overall, this analysis suggests that the data relating to fatigue is adequately complete and stable to use.

## 3.7 Chapter summary

In this chapter I discussed the methodological issues inherent to risk studies in symptomatic cohorts, including how to choose appropriate comparisons and statistical methods that fit the study purpose. It also represents the first in-depth assessment of issues specifically arising from conducting such studies using electronic health records (EHRs), which are not primarily intended for research. Issues include overly granular or 'messy' coding systems, missing or selectively recorded data items, loss to follow up as patients move practices, and difficulties clearly defining the state of patients (e.g. having multiple co-occurring symptoms, or being currently disease-free). I explained the approaches I used in my empirical studies describing the risk of cancer and other diseases in patients with new-onset fatigue. This practical information will support other researchers to develop future similar studies of disease risk in symptomatic cohorts using EHRs.

# 4. Chapter 4: Risk of cancer following primary care presentation with fatigue: A population-based cohort study of a quarter of a million patients

## 4.1 Chapter rationale

This chapter aimed to establish the risk of undetected cancer among patients with fatigue, in comparison to the general population, and the full range of cancer sites associated with fatigue. This would establish 'baseline' estimates of risk, and guide the development of subsequent chapters, which would seek to modify risk estimates with the inclusion of additional features (e.g. other presenting symptoms, diagnostic test results). This chapter also examined the period of time after presentation patients were at increased risk, to establish the appropriate follow up period for subsequent studies.

## 4.2 Publication

This chapter has been published in the peer reviewed journal, British Journal of Cancer:

White, B., Rafiq, M., Gonzalez-Izquierdo, A., Hamilton, W., Price, S., & Lyratzopoulos, G. (2022). Risk of cancer following primary care presentation with fatigue: a population-based cohort study of a quarter of a million patients. British Journal of Cancer, 126(11), 1627–1636. https://doi.org/10.1038/s41416-022-01733-6



## 4.3 Author contributions

Authors: Becky White, Meena Rafiq, Arturo Gonzalez-Izquierdo, Willie Hamilton, Sarah Price, Georgios Lyratzopoulos

BW, GL, and MR conceived and designed the study. BW and AGI managed and BW analysed the data, under the supervision of GL, MR, and AGI. MR and GL provided clinical input, and WH and SP developed medical code lists used for case identification and advised on the presentation and discussion of results. All authors contributed to drafting and revising the article.

## 4.4 Abstract

**Background**: The management of adults presenting with fatigue presents a diagnostic challenge, particularly regarding possible underlying cancer.

**Methods:** Using electronic health records, I examined cancer risk in patients presenting to primary care with new-onset fatigue in England during 2007-2013, compared to general population estimates. I examined variation by age, sex, deprivation, and time following presentation.

**Findings**: Of 250,606 patients presenting with fatigue, 12-month cancer risk exceeded 3% in men aged 65 and over and women aged 80 and over, and 6% in men aged 80 and over. Nearly half (47%) of cancers were diagnosed within three months from first fatigue presentation. Site-specific cancer risk was higher than the general population for most cancers studied, with greatest relative increases for leukaemia, pancreatic and brain cancers.

**Conclusions**: In older patients, new-onset fatigue is associated with cancer risk exceeding current thresholds for urgent specialist investigation. Future research should consider how risk is modified by the presence or absence of other signs and symptoms. Excess cancer risk wanes rapidly after three months, which could inform the duration of a 'safety-netting' period. Fatigue presentation is not strongly predictive of any single cancer, although certain cancers are over-represented; this knowledge can help prioritise diagnostic strategies.

## 4.5 Background

Many cancer patients are diagnosed after presenting to a general practitioner with non-site specific symptoms of relatively low specificity(17), for which there are limited referral or investigation guidelines(13). Cancer patients least likely to be diagnosed following fast-track referral are those with cancers typically characterised by such non-site specific symptoms (e.g. fatigue, weight loss etc.), which have low positive predictive value (PPV) for any single cancer(12). Consequently, these patients often experience prolonged intervals before diagnosis(13).

Fatigue is a relatively common presenting symptom in primary care, being the principal complaint in an estimated 5-7% of consultations(21–24), and even more common in the general population, with 15-40% of people reporting experiencing fatigue in the last two or four weeks(25,26). Fatigue is known to be a presenting feature of several cancers(28–31). Diagnostic guidelines by the National Institute for Health and Care Excellence (NICE) regarding fatigue recommend urgent two-week-wait referral only for specific presentations where available evidence shows the positive predictive value (PPV) for specific cancer diagnosis (usually within 12 months) exceeds 3%(28,29,43). However, the range of cancer sites associated with fatigue and their relative specific risk is not adequately described in current literature, which is dominated by studies focusing on individual cancer sites. Nonetheless, the limited available evidence suggests that the predictive value of fatigue as a single presenting symptom of colorectal, lung, urological cancer and leukaemia appears to be low(29,32,35,38). As a relatively common symptom, fatigue can also signal a range of other conditions, including self-limiting illnesses (e.g. short-term post-viral fatigue); depression; chronic fatigue syndrome; a range of other causes (e.g. hypothyroidism, vitamin deficiency, iron deficiency, coeliac disease etc.); and more rarely, autoimmune disease such as lupus or chronic infection such as hepatitis C(22,39–42).

Given the low PPV of fatigue for cancer, and the range of possible other causes, primary and secondary care clinicians must assess which patients presenting with fatigue are more likely to have cancer, thereby requiring investigation and specialist assessment. Consequently, investigating the predictive value of fatigue for any cancer and specific types of cancer, for different age and sex groups, is important to help determine appropriate diagnostic strategies to diagnose or rule out specific cancers efficiently. Although patients who seek medical help for fatigue are not representative of the broader population of individuals with fatigue in the community(25,26,147), understanding their cancer risk when they first present to primary care is important to support general practitioners' decisions about their management. It is also unclear how long patients who present with new-onset fatigue remain at greater risk of being diagnosed with cancer after initial presentation, and therefore how long healthcare professionals and patients should be alert to changing symptoms and other diagnostic clues (i.e. the 'safety-netting' period)(148).

Therefore, this study aimed to establish the risk of present but as-yet-undetected cancer (overall and by specific cancer site) among patients who present with 'new onset' fatigue to a general practitioner, and related changes over time in such risks in the months after initial presentation. It also aimed to contextualise the excess risk in these patients through a comparison with cancer risk in the general population for persons of the same sex and age band.

## 4.6 Methods

### 4.6.1 Study design and data source

I conducted a cohort study of patients with a record of fatigue presentation in primary care in England between 2007 and 2013, using electronic health records (EHRs) from the Clinical Practice Research Datalink (CPRD) GOLD (March 2019 database build Index of Multiple Deprivation (IMD) quintile, and cancers diagnosed from 2006-2015 in National Cancer Registration and Analysis Service (NCRAS) data. For more information about CPRD and linked datasets, see Section 3.6.2.

### 4.6.2 Cohort identification

In step 1, patients were included in the study if they had a code for fatigue recorded during a consultation in CPRD within the overall study period (2007-2013). In step 2, patients were included if at least one of their fatigue records was 'eligible', i.e. the date occurred *after* all of the following events: the date the patient's practice was up to standard regarding research quality, the patient was registered to the practice for at least a year, and the patient was 30 years old. The date also had to occur *before* all of the following events (if relevant): the date the practice last submitted data to CPRD, the patient left the practice, the patient was aged 100 years or over, or the patient's death.

To produce risk estimates relevant to primary care clinicians, I aimed to ensure the study population broadly represented patients attending primary care with *new-onset* fatigue, to minimise the likelihood that it was attributable to a previously diagnosed condition or disease (including cancer). Therefore, in step 3, I excluded a small group of patients who had an 'ineligible' record of fatigue in the year before their first eligible fatigue record (as a patient could have had a prior record of fatigue before the date they entered the study as defined in step 2 (e.g. before the patient was 30 years old)). This meant that patients did not enter the study midway through a series of consultations for fatigue. However, if such patients had another eligible fatigue record more than a year later, the later record was selected and the patient was included.

To minimise the likelihood that fatigue was attributable to a previous cancer diagnosis, in step 4, I also excluded patients if there was a cancer diagnosis recorded in NCRAS in the year before or on the same day as their first eligible fatigue record. In these cases, the patient was still included if they had another eligible fatigue record more than a year after the first eligible fatigue record. I conducted a sensitivity analysis where I extended the look-back period in steps 3 and 4 to two years, and also where these two exclusions were not conducted. In Results, Figure 4.1 illustrates steps 1 to 4.

According to National Institute for Health and Care Excellence (NICE) Guidelines(54), there is no universal definition of fatigue. Therefore, in this study, fatigue was defined by a list of Read codes collated using methods developed by WH and refined by SP(94) (Supplementary Appendix 10.4.2). Although the study was concerned with new onset fatigue, I included codes for Chronic Fatigue Syndrome (CFS) and Post Viral Fatigue Syndrome (PVFS), because some patients with fatigue could initially be misdiagnosed with CFS or PVFS. This is analogous to previous research which has highlighted misdiagnoses of Irritable Bowel Syndrome (IBS) in some patients with colorectal cancer(70,149). Nonetheless, I conducted a sensitivity analysis to ascertain whether excluding CFS

and PVFS impacted cancer risk estimates. I did not consider other pre-existing conditions (e.g. anaemia) that could have explained the presence of fatigue.

### 4.6.3 Follow up and outcomes

Follow up began with the patient's first eligible record of fatigue during the study period (termed the 'index' record). Follow up ended either at one year following index record, or the first cancer diagnosis, if earlier. As NCRAS data was used to define the outcome, patients could remain in the study even after they had left their GP practice or their practice had exited CPRD. After follow up ended, patients could not re-enter the study with a subsequent fatigue record (i.e. patients were included in the study once).

The main outcome was diagnosis of cancer recorded in cancer registry (NCRAS) data within 12 months after first (index) fatigue record. One year was chosen to enable comparison with most primary research underpinning diagnostic guidelines regarding fatigue (NICE)(28,29,43). I conducted a supplementary analysis following patients up to two years, which confirmed that one year was long enough to capture relevant cancer cases (Supplementary Appendix 10.4.3). Cancers included any malignant neoplasms, excluding non-melanoma skin cancer (International Classification of Diseases 10th Revision (ICD-10) codes C00-C99 excl. C45). Benign brain tumours were not included. Cancer site definitions were adapted from previously published ICD-10 codes(150). Rarer cancers were combined into anatomically related groups or, where this was not possible, grouped into 'other cancers' (Supplementary Appendix 10.4.4).

### 4.6.4 Statistical analysis

The age, sex, and deprivation quintile of patients with fatigue were compared to the general population in England (Table 4.1, Supplementary Appendix 10.4.6). I calculated the risk of cancer, for all cancers combined and stratified by cancer site diagnosed. Analyses were stratified by sex and five-year age band, as previous research has identified substantial variability in cancer incidence between these groups(107). I aimed to estimate values to a level of precision where 95% confidence intervals were no wider than 0.5 percentage points either side of the cancer risk estimates. Assuming proportions of 3%, I calculated that sample sizes of at least 3,700 patients were needed in each age-sex strata.

For Table 4.2, Figure 4.2 and Figure 4.3, I calculated absolute and relative differences in cancer risk between patients with fatigue and the general population (derived using incident cancer registration statistics for England in 2011(151) and corresponding mid-year population estimates)(152), for each age-sex stratum. For calculations using the general population estimates, I assumed that no person was diagnosed with more than one cancer during a year. I conducted a separate supplementary analysis of cancer risk by deprivation quintile, as there were no directly comparable general population cancer risk estimates that would allow me to also adjust for age and sex.

For secondary analyses, I also used general population estimates to derive expected cancer risk for all persons with fatigue combined, as described in prior literature(153–155). I directly standardised general population estimates by multiplying the total number of patients in each sex and five-year

age band in the fatigue cohort by the corresponding annual cancer incidence in the general population, thereby obtaining the expected age- and sex-specific number of incident cancers in the fatigue cohort. These were summed to calculate expected cancers for men, women, and both sexes combined.

For Table 4.3, I anticipated that risk estimates would have adequate precision for the comparison between observed and expected risk for certain cancer sites, though estimates for particularly rare cancer sites (under 30 cases) were not shown.

To better describe variability in excess cancer risk after the initial record of fatigue, in Figure 4.4 and Supplementary Appendix 10.4.3, I compared the observed and expected number of cancer cases by month of follow-up. To derive expected monthly cases, annual cancer incidence in the general population was divided by 12 and then age- and sex-standardised to derive expected monthly cases. I subtracted the expected from the observed monthly cases, to calculate excess cases each month.

Data management was conducted in MySQL Workbench version 6.1, with all statistical analysis conducted in Stata version 16. Age and sex standardisation was performed using the user-written *distrate* command for Stata(156), with 95% confidence intervals calculated using the Dobson et al. method for rare outcomes(157). Pearson's chi-square tests (which were robust to assumptions about data distribution and degree of homoscedasticity) were used to assess statistical significance of differences in cancer incidence between the fatigue cohort and the general population. P values < 0.05 were considered significant. I used Strengthening the Reporting of Observational studies in Epidemiology (STROBE) guidelines for cohort studies(158) to report this study (Supplementary Appendix 10.4.5).

## 4.7 Findings

### 4.7.1 Cohort description

Of the 278,821 individuals who had a record of fatigue in primary care between 2007 and 2013, 250,606 (90%) had at least one 'eligible' record within the patient's inclusion period, without a cancer diagnosis or an 'ineligible' fatigue record in the previous year (Figure 4.1). These were included in the study cohort.

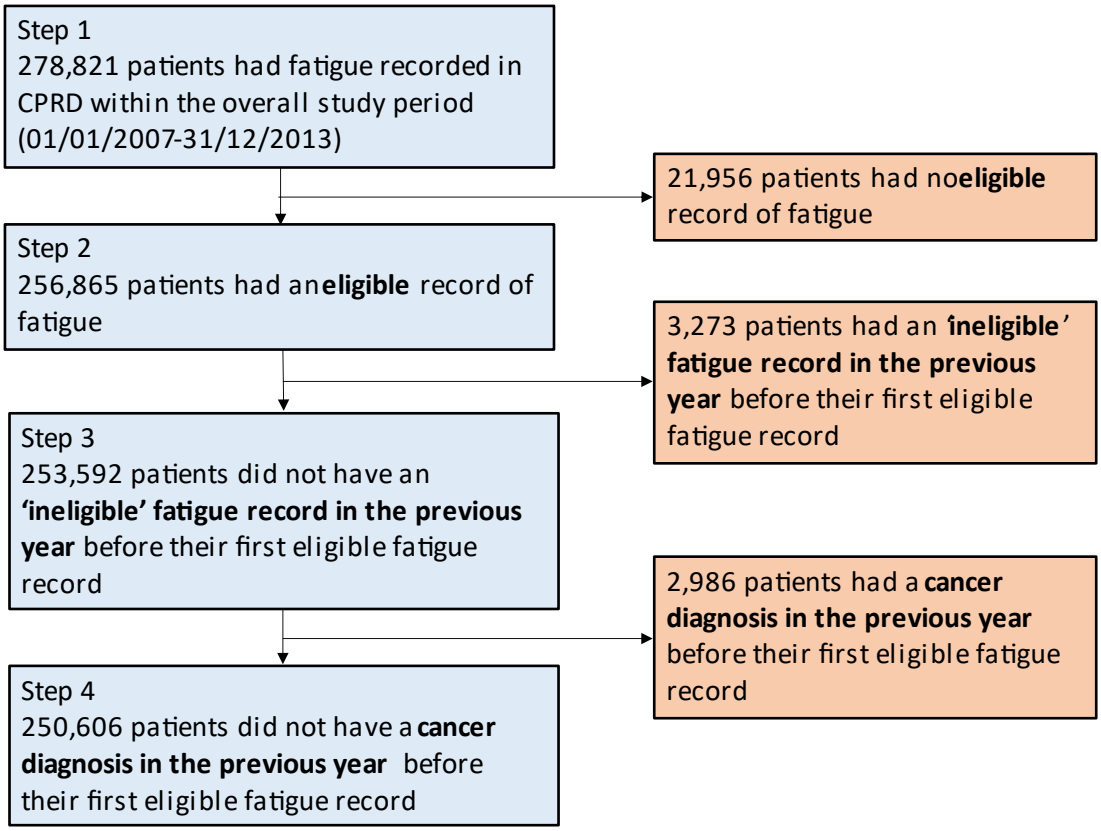



*Figure 4.1. Study inclusions and exclusions*

There was a preponderance of women in the cohort (68%), compared to 52% in the general population(152). The study cohort was also slightly older than the general population. For example, of patients with cancer, 49% of men and 45% of women with fatigue were aged 75 years and over, compared to 37% and 37% in the general population, respectively (Table 4.1). The study cohort was also slightly less deprived, with 23% in the least deprived quintile compared to 18% of people aged 30 years and over in the general population (Supplementary Appendix 10.4.6). Regarding specific subcodes, 0.81% (n = 2,033) of the study cohort had a first fatigue record for either Chronic Fatigue Syndrome (CFS) or Post Viral Fatigue Syndrome (PVFS) (Supplementary Appendix 10.4.8).

*Table 4.1. Gender and age characteristics of patients presenting to primary care with fatigue compared to population estimates, by subsequent cancer diagnosis within a year after first presentation*

| Age group[a] | Patients with fatigue | | | | England population | | | |
|---|---|---|---|---|---|---|---|---|
| | No cancer | | Cancer[b] | | No cancer | | Cancer[c] | |
| | n | % | n | % | n | % | n | % |
| Men | | | | | | | | |
| 30-34 | 4,737 | 5.98 | <5[d] | - | 1,764,208 | 11.13 | 1,074 | 0.79 |
| 35-39 | 6,194 | 7.83 | 12 | 0.60 | 1,754,358 | 11.07 | 1,299 | 0.95 |
| 40-44 | 7,600 | 9.60 | 19 | 0.96 | 1,919,735 | 12.11 | 2,183 | 1.60 |
| 45-49 | 8,147 | 10.29 | 30 | 1.51 | 1,922,296 | 12.13 | 3,804 | 2.78 |
| 50-54 | 7,960 | 10.06 | 80 | 4.03 | 1,692,822 | 10.68 | 6,166 | 4.51 |
| 55-59 | 7,864 | 9.93 | 117 | 5.89 | 1,474,698 | 9.30 | 10,086 | 7.38 |
| 60-64 | 8,553 | 10.81 | 208 | 10.47 | 1,534,388 | 9.68 | 17,610 | 12.88 |
| 65-69 | 6,742 | 8.52 | 254 | 12.78 | 1,221,416 | 7.71 | 21,197 | 15.51 |
| 70-74 | 6,270 | 7.92 | 298 | 15.00 | 941,209 | 5.94 | 22,369 | 16.36 |
| 75-79 | 6,040 | 7.63 | 362 | 18.22 | 739,908 | 4.67 | 21,219 | 15.52 |
| 80-84 | 4,744 | 5.99 | 328 | 16.51 | 507,744 | 3.20 | 16,448 | 12.03 |
| 85+ | 4,305 | 5.44 | 275 | 13.84 | 376,656 | 2.38 | 13,253 | 9.69 |
| *Mean, median age* | *58, 58* | | *73, 74* | | *54, 52* | | *70, 71* | |
| Total men | 79,156 | | 1,987 | | 15,849,438 | | 136,708 | |
| Women | | | | | | | | |
| 30-34 | 16,215 | 9.69 | 22 | 1.05 | 1,762,200 | 11.13 | 1,668 | 0.79 |
| 35-39 | 18,706 | 11.18 | 32 | 1.52 | 1,764,022 | 11.07 | 2,693 | 0.95 |
| 40-44 | 20,142 | 12.03 | 59 | 2.81 | 1,954,742 | 12.11 | 4,823 | 1.60 |
| 45-49 | 19,866 | 11.87 | 99 | 4.71 | 1,958,271 | 12.13 | 7,705 | 2.78 |
| 50-54 | 16,628 | 9.94 | 119 | 5.67 | 1,713,902 | 10.68 | 9,689 | 4.51 |
| 55-59 | 13,851 | 8.28 | 137 | 6.52 | 1,508,030 | 9.30 | 10,687 | 7.38 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 60-64 | 12,591 | 7.52 | 219 | 10.43 | 1,594,525 | 9.68 | 15,768 | 12.88 |
| 65-69 | 10,523 | 6.29 | 231 | 11.00 | 1,298,592 | 7.71 | 16,415 | 15.51 |
| 70-74 | 10,146 | 6.06 | 238 | 11.33 | 1,055,292 | 5.94 | 15,681 | 16.36 |
| 75-79 | 10,432 | 6.23 | 316 | 15.05 | 901,399 | 4.67 | 15,805 | 15.52 |
| 80-84 | 8,654 | 5.17 | 307 | 14.62 | 726,834 | 3.20 | 14,821 | 12.03 |
| 85+ | 9,609 | 5.74 | 321 | 15.29 | 786,631 | 2.38 | 16,778 | 9.69 |
| *Mean, median age* | *55, 52* | | *70, 72* | | *55, 53* | | *68, 69* | |
| Total women | 167,363 | | 2,100 | | 17,024,440 | | 132,533 | |

*aAge at first presentation. Mean and median ages for available population estimates were estimated from aggregated five-year age bands. bCancer diagnoses between 2007-2014, 12 months after first presentation with fatigue to primary care in 2007-2013, while aged 30-99 years. cEstimated 12-month population incidence, based on annual number of cancer diagnoses and mid-year population estimates for England, 2011. Available population estimates include patients aged > 99 years. This was estimated to account for < 0.9% of people aged 85+ in this analysis, thus would have a negligible impact on cancer incidence estimates for this age group. dCell counts under 5 are suppressed to reduce statistical disclosure risk.*

### 4.7.2 Risk of cancer

For men, the risk of any cancer diagnosis within a year after the first fatigue record ranged from below 1% in each five-year age band from those aged 30-49, to 3-6% in age bands between 65-79 years, and over 6% for those aged 80 years and over (Figure 4.2, Table 4.2). Cancer risk was higher in men with fatigue than men in the general population in every age band from 35 years and over ($p < 0.01$ for all), and was typically at least twice as high, with no clear trend by age.

The risk of cancer in women with fatigue ranged from below 1% in those aged 30 to 59 years, to over 3% in those aged 80 years and over (Figure 4.3, Table 4.2). Cancer risk was higher in women with fatigue than women in the general population in every age band from 45 years and over ($p < 0.05$ for all), rising to between 55% and 75% higher among those aged 60 years and over.

Comparing patterns in men and women, relative increases in cancer risk compared to the general population appeared higher in men than women, although differences in each age band were not generally statistically significant (see confidence intervals for risk ratios in Table 4.2). In supplementary analysis, cancer risk was similar across deprivation quintiles (Supplementary Appendix 10.4.7).

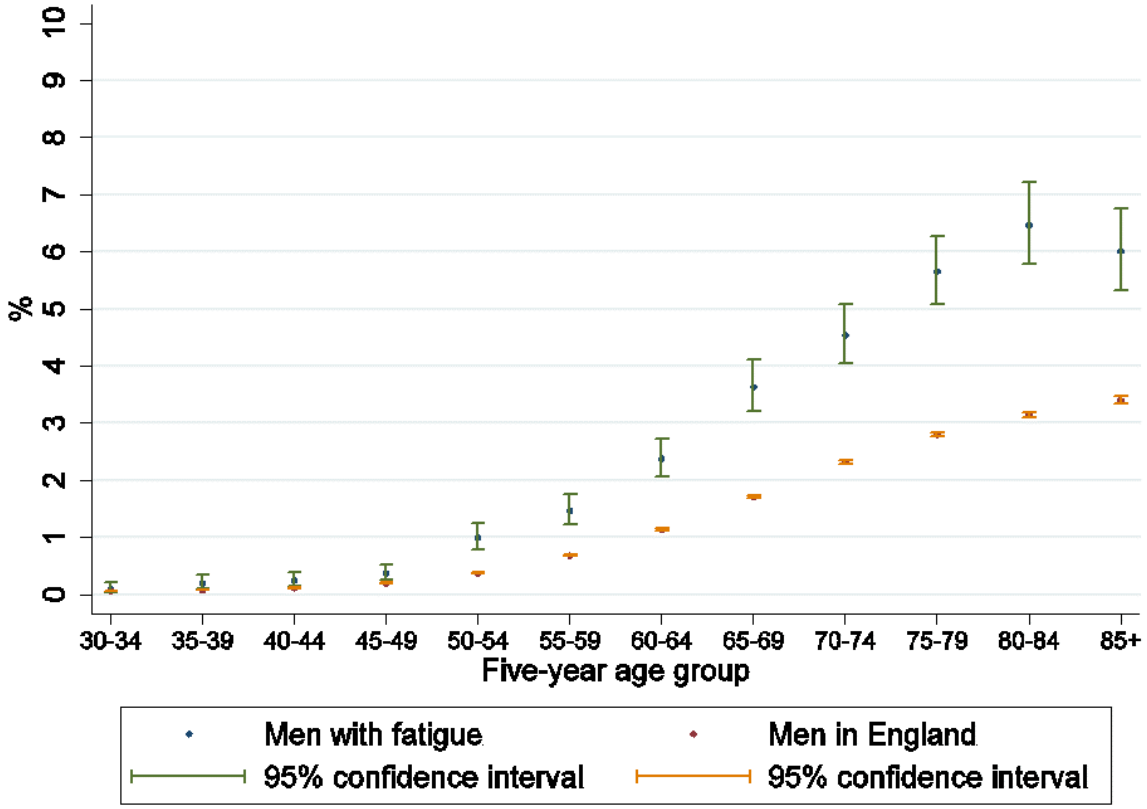



*Figure 4.2. Cancer risk (%) within a year for men with fatigue, compared to men in England.*

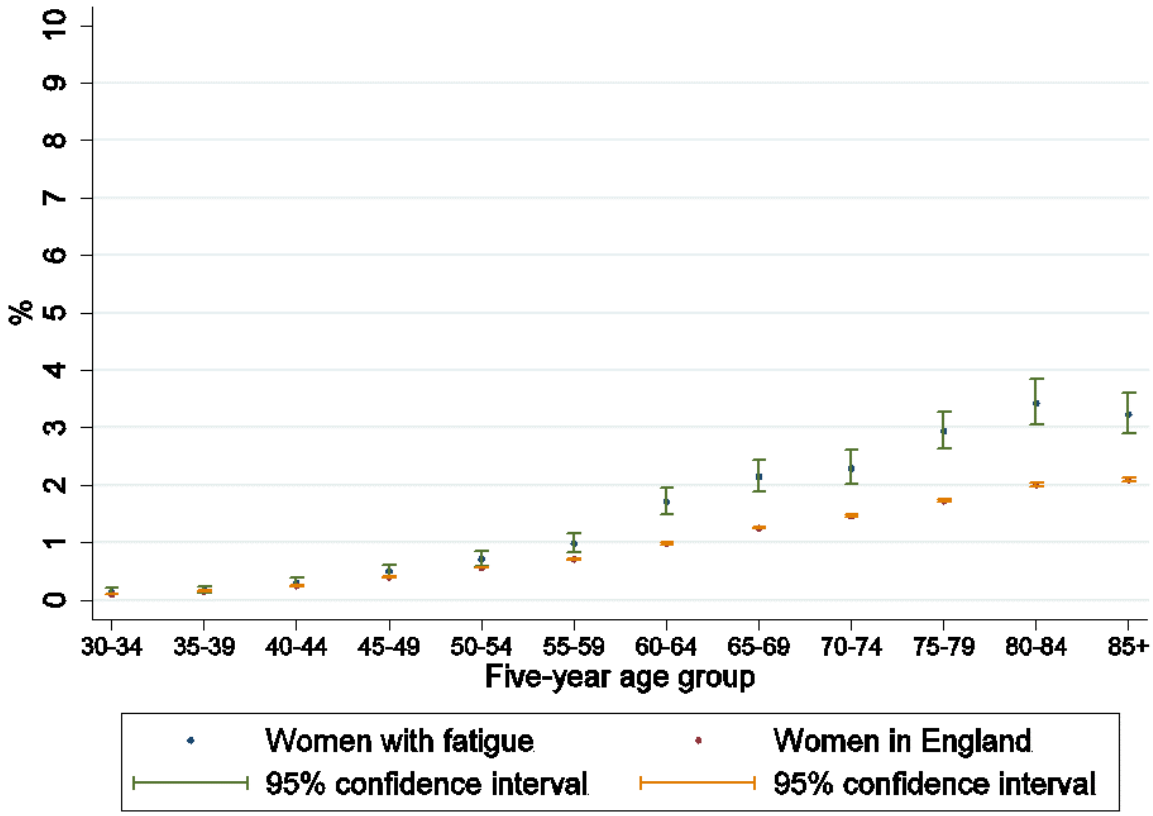



*Figure 4.3. Cancer risk (%) within a year for women with fatigue, compared to women in England.*

*Table 4.2. Number and proportion of patients diagnosed with cancer within a year after presenting to primary care with fatigue compared to population estimates, by gender and age band*

| | Patients with fatigue | | | England population | | | Absolute difference (%) | Risk ratio [lci,uci] | P-value |
|---|---|---|---|---|---|---|---|---|---|
| | Cancer[a] | | Total | Cancer[b] | | Total | | | |
| | n | % [lci,uci] | N | n | % [lci,uci] | N | | | |
| Men | | | | | | | | | |
| 30-34 | <5[c] | - | 4,741 | 1,074 | 0.06 [0.06,0.06] | 1,765,282 | - | - | - |
| 35-39 | 12 | 0.19 [0.1,0.34] | 6,206 | 1,299 | 0.07 [0.07,0.08] | 1,755,657 | 0.12 | 2.61 [1.35,4.58] | 0.001 |
| 40-44 | 19 | 0.25 [0.15,0.39] | 7,619 | 2,183 | 0.11 [0.11,0.12] | 1,921,918 | 0.14 | 2.2 [1.32,3.44] | 0.001 |
| 45-49 | 30 | 0.37 [0.25,0.52] | 8,177 | 3,804 | 0.2 [0.19,0.2] | 1,926,100 | 0.17 | 1.86 [1.25,2.66] | 0.001 |
| 50-54 | 80 | 1 [0.79,1.24] | 8,040 | 6,166 | 0.36 [0.35,0.37] | 1,698,988 | 0.63 | 2.74 [2.17,3.42] | <0.001 |
| 55-59 | 117 | 1.47 [1.21,1.76] | 7,981 | 10,086 | 0.68 [0.67,0.69] | 1,484,784 | 0.79 | 2.16 [1.78,2.59] | <0.001 |
| 60-64 | 208 | 2.37 [2.06,2.72] | 8,761 | 17,610 | 1.13 [1.12,1.15] | 1,551,998 | 1.24 | 2.09 [1.82,2.4] | <0.001 |
| 65-69 | 254 | 3.63 [3.2,4.11] | 6,996 | 21,197 | 1.71 [1.68,1.73] | 1,242,613 | 1.92 | 2.13 [1.87,2.41] | <0.001 |
| 70-74 | 298 | 4.54 [4.04,5.08] | 6,568 | 22,369 | 2.32 [2.29,2.35] | 963,578 | 2.22 | 1.95 [1.74,2.19] | <0.001 |
| 75-79 | 362 | 5.65 [5.09,6.27] | 6,402 | 21,219 | 2.79 [2.75,2.83] | 761,127 | 2.87 | 2.03 [1.82,2.25] | <0.001 |
| 80-84 | 328 | 6.47 [5.79,7.21] | 5,072 | 16,448 | 3.14 [3.09,3.19] | 524,192 | 3.33 | 2.06 [1.84,2.3] | <0.001 |
| 85+ | 275 | 6 [5.32,6.76] | 4,580 | 13,253 | 3.4 [3.34,3.46] | 389,909 | 2.61 | 1.77 [1.56,1.99] | <0.001 |
| Women | | | | | | | | | |
| 30-34 | 22 | 0.14 [0.08,0.21] | 16,237 | 1,668 | 0.09 [0.09,0.1] | 1,763,868 | 0.04 | 1.43 [0.9,2.18] | 0.092 |
| 35-39 | 32 | 0.17 [0.12,0.24] | 18,738 | 2,693 | 0.15 [0.15,0.16] | 1,766,715 | 0.02 | 1.12 [0.76,1.59] | 0.522 |
| 40-44 | 59 | 0.29 [0.22,0.38] | 20,201 | 4,823 | 0.25 [0.24,0.25] | 1,959,565 | 0.05 | 1.19 [0.9,1.53] | 0.190 |
| 45-49 | 99 | 0.5 [0.4,0.6] | 19,965 | 7,705 | 0.39 [0.38,0.4] | 1,965,976 | 0.10 | 1.27 [1.03,1.54] | 0.020 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 50-54 | 119 | 0.71 [0.59,0.85] | 16,747 | 9,689 | 0.56 [0.55,0.57] | 1,723,591 | 0.15 | 1.26 [1.05,1.51] | 0.011 |
| 55-59 | 137 | 0.98 [0.82,1.16] | 13,988 | 10,687 | 0.7 [0.69,0.72] | 1,518,717 | 0.28 | 1.39 [1.17,1.65] | <0.001 |
| 60-64 | 219 | 1.71 [1.49,1.95] | 12,810 | 15,768 | 0.98 [0.96,0.99] | 1,610,293 | 0.73 | 1.75 [1.52,2] | <0.001 |
| 65-69 | 231 | 2.15 [1.88,2.44] | 10,754 | 16,415 | 1.25 [1.23,1.27] | 1,315,007 | 0.90 | 1.72 [1.5,1.96] | <0.001 |
| 70-74 | 238 | 2.29 [2.01,2.6] | 10,384 | 15,681 | 1.46 [1.44,1.49] | 1,070,973 | 0.83 | 1.57 [1.37,1.78] | <0.001 |
| 75-79 | 316 | 2.94 [2.62,3.28] | 10,748 | 15,805 | 1.72 [1.7,1.75] | 917,204 | 1.22 | 1.71 [1.52,1.91] | <0.001 |
| 80-84 | 307 | 3.43 [3.05,3.83] | 8,961 | 14,821 | 2 [1.97,2.03] | 741,655 | 1.43 | 1.71 [1.53,1.92] | <0.001 |
| 85+ | 321 | 3.23 [2.89,3.61] | 9,930 | 16,778 | 2.09 [2.06,2.12] | 803,409 | 1.14 | 1.55 [1.38,1.73] | <0.001 |

*aCancer diagnoses between 2007-2014, 12 months after first presentation with fatigue to primary care in 2007-2013, while aged 30-99 years. bEstimated 12-month population incidence, based on annual number of cancer diagnoses and mid-year population estimates for England, 2011. Available population estimates include patients aged > 99 years. This was estimated to account for < 0.9% of people aged 85+ in this analysis, thus would have a negligible impact on cancer incidence estimates for this age group. cCell counts under 5 are suppressed to reduce statistical disclosure risk.*

### 4.7.3 Frequency of specific cancer sites

For men, site-specific cancer risk was higher than expected for 13 of the 16 cancer sites studied (all p values < 0.001). Although their absolute associated risk was low (≤ 0.12%), in relative terms the risk of diagnosis of leukaemia, pancreatic, and brain cancers was 3- to 4-fold greater than expected ($p < 0.001$). The overall case mix of cancer sites was different to expected ($p < 0.001$), although the three most common cancers in men in the general population (prostate, lung and colorectal) still accounted for the majority (52% (n = 1,041)) of observed cases in my sample (Table 4.3).

For women, site-specific cancer risk was higher than expected for 13 of the 17 cancer sites studied (all p values < 0.02). Although their absolute associated risk was low (≤ 0.06%), in relative terms, the risk of diagnosis of leukaemia, pancreatic, and brain cancers was 2- to 4-fold greater than expected ($p < 0.001$). The overall case mix of cancers was different to expected ($p < 0.001$), although the three most common cancers in women (breast, lung and colorectal cancers) in the general population still accounted for half (50%, n = 1,055) of observed cases in my sample (Table 4.3).

*Table 4.3. First cancer site diagnosed within a year, as a proportion of patients presenting to primary care with fatigue, observed compared to expected*

| | Observed[a] | | Expected[b] | | Absolute difference (%) | Risk ratio [lci,uci] | P-value[c] |
|---|---|---|---|---|---|---|---|
| | n | % [lci,uci] | n | % [lci,uci] | | | |
| Men | | | | | | | |
| Chi2 (P-value) comparing observed and expected distribution of cancer sites: <0.001 | | | | | | | |
| All cancers | 1,987 | 2.45 [2.34,2.56] | 980 | 1.21 [1.2,1.21] | 1.24 | 2.03 [1.94,2.12] | <0.001 |
| Prostate | 406 | 0.5 [0.45,0.55] | 255 | 0.31 [0.31,0.32] | 0.19 | 1.59 [1.44,1.75] | <0.001 |
| Lung and mesothelioma | 384 | 0.47 [0.43,0.52] | 159 | 0.2 [0.19,0.2] | 0.28 | 2.41 [2.18,2.67] | <0.001 |
| Colorectal | 251 | 0.31 [0.27,0.35] | 138 | 0.17 [0.17,0.17] | 0.14 | 1.82 [1.61,2.06] | <0.001 |
| Upper gastro-intestinal | 125 | 0.15 [0.13,0.18] | 62 | 0.08 [0.07,0.08] | 0.08 | 2.03 [1.7,2.42] | <0.001 |
| Lymphoma | 121 | 0.15 [0.12,0.18] | 42 | 0.05 [0.05,0.05] | 0.10 | 2.89 [2.41,3.45] | <0.001 |
| Leukaemia | 98 | 0.12 [0.1,0.15] | 28 | 0.03 [0.03,0.04] | 0.09 | 3.49 [2.86,4.27] | <0.001 |
| Unknown primary | 83 | 0.1 [0.08,0.13] | 29 | 0.04 [0.03,0.04] | 0.07 | 2.87 [2.31,3.57] | <0.001 |
| Pancreas | 81 | 0.1 [0.08,0.12] | 27 | 0.03 [0.03,0.03] | 0.07 | 3.03 [2.43,3.78] | <0.001 |
| Kidney | 71 | 0.09 [0.07,0.11] | 29 | 0.04 [0.03,0.04] | 0.05 | 2.43 [1.92,3.07] | <0.001 |
| Other malignant neoplasms | 65 | 0.08 [0.06,0.1] | 28 | 0.04 [0.03,0.04] | 0.04 | 2.28 [1.79,2.92] | <0.001 |
| Bladder | 56 | 0.07 [0.05,0.09] | 50 | 0.06 [0.06,0.06] | 0.01 | 1.11 [0.85,1.44] | 0.438 |
| Brain and other CNS | 55 | 0.07 [0.05,0.09] | 13 | 0.02 [0.02,0.02] | 0.05 | 4.25 [3.25,5.55] | <0.001 |
| Melanoma | 43 | 0.05 [0.04,0.07] | 34 | 0.04 [0.04,0.04] | 0.01 | 1.25 [0.93,1.69] | 0.146 |
| Multiple myeloma | 41 | 0.05 [0.04,0.07] | 17 | 0.02 [0.02,0.02] | 0.03 | 2.48 [1.82,3.38] | <0.001 |
| Liver | 40 | 0.05 [0.04,0.07] | 16 | 0.02 [0.02,0.02] | 0.03 | 2.5 [1.83,3.42] | <0.001 |
| Head and neck | 39 | 0.05 [0.03,0.07] | 35 | 0.04 [0.04,0.04] | 0.00 | 1.11 [0.81,1.52] | 0.523 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Sarcoma | < 30 | - | - | - | - | - | - |
| Thyroid | < 30 | - | - | - | - | - | - |
| Testis | < 30 | - | - | - | - | - | - |
| Breast | < 30 | - | - | - | - | - | - |
| Total men | 81,143 | | 81,143 | | | | |

Women

Chi2 (P-value) comparing observed and expected distribution of cancer sites: <0.001

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| All cancers | 2,100 | 1.24 [1.19,1.29] | 1,348 | 0.8 [0.79,0.8] | 0.44 | 1.56 [1.49,1.63] | <0.001 |
| Breast | 426 | 0.25 [0.23,0.28] | 409 | 0.24 [0.24,0.24] | 0.01 | 1.04 [0.95,1.15] | 0.408 |
| Lung and mesothelioma | 328 | 0.19 [0.17,0.22] | 165 | 0.1 [0.1,0.1] | 0.10 | 1.99 [1.78,2.22] | <0.001 |
| Colorectal | 301 | 0.18 [0.16,0.2] | 157 | 0.09 [0.09,0.09] | 0.09 | 1.92 [1.71,2.15] | <0.001 |
| Lymphoma | 112 | 0.07 [0.05,0.08] | 53 | 0.03 [0.03,0.03] | 0.03 | 2.1 [1.75,2.54] | <0.001 |
| Pancreas | 105 | 0.06 [0.05,0.08] | 39 | 0.02 [0.02,0.02] | 0.04 | 2.71 [2.23,3.29] | <0.001 |
| Unknown primary | 101 | 0.06 [0.05,0.07] | 47 | 0.03 [0.03,0.03] | 0.03 | 2.13 [1.75,2.6] | <0.001 |
| Ovary | 95 | 0.06 [0.05,0.07] | 57 | 0.03 [0.03,0.03] | 0.02 | 1.65 [1.35,2.02] | <0.001 |
| Uterus | 77 | 0.05 [0.04,0.06] | 69 | 0.04 [0.04,0.04] | 0.00 | 1.12 [0.9,1.41] | 0.313 |
| Brain and other CNS | 66 | 0.04 [0.03,0.05] | 17 | 0.01 [0.01,0.01] | 0.03 | 3.97 [3.11,5.08] | <0.001 |
| Leukaemia | 65 | 0.04 [0.03,0.05] | 29 | 0.02 [0.02,0.02] | 0.02 | 2.23 [1.75,2.86] | <0.001 |
| Melanoma | 65 | 0.04 [0.03,0.05] | 54 | 0.03 [0.03,0.03] | 0.01 | 1.21 [0.94,1.54] | 0.133 |
| Upper gastro-intestinal | 64 | 0.04 [0.03,0.05] | 45 | 0.03 [0.03,0.03] | 0.01 | 1.43 [1.12,1.83] | 0.004 |
| Kidney | 56 | 0.03 [0.02,0.04] | 27 | 0.02 [0.02,0.02] | 0.02 | 2.06 [1.58,2.68] | <0.001 |
| Other malignant neoplasms | 56 | 0.03 [0.02,0.04] | 39 | 0.02 [0.02,0.02] | 0.01 | 1.43 [1.1,1.86] | 0.008 |
| Bladder | 31 | 0.02 [0.01,0.03] | 27 | 0.02 [0.02,0.02] | 0.00 | 1.15 [0.81,1.64] | 0.447 |
| Liver | 30 | 0.02 [0.01,0.03] | 14 | 0.01 [0.01,0.01] | 0.01 | 2.21 [1.54,3.18] | <0.001 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Multiple myeloma | 30 | 0.02 [0.01,0.03] | 18 | 0.01 [0.01,0.01] | 0.01 | 1.62 [1.13,2.33] | 0.008 |
| Cervix | <30 | - | - | - | - | - | - |
| Thyroid | <30 | - | - | - | - | - | - |
| Sarcoma | <30 | - | - | - | - | - | - |
| Head and neck | <30 | - | - | - | - | - | - |
| Vulva | <30 | - | - | - | - | - | - |
| Total women | 169,463 | | 169,463 | | | | |

*[a]Cancer diagnoses between 2007-2014, 12 months after first presentation with fatigue to primary care in 2007-2013. bExpected cases for the age distribution of men and women with fatigue, based on five-year age band and sex-specific estimated monthly population incidence, using annual number of cancer diagnoses and mid-year population estimates for England, 2011. Results not shown for cancers with fewer than 30 observed cases. [c]Pearson's chi-square tests (which were robust to assumptions about data distribution and degree of homoscedasticity) were used to assess statistical significance of differences in cancer incidence between the fatigue cohort and the general population. P values < 0.05 were considered significant.*

### 4.7.4 Distribution of incident cases by month following fatigue presentation

Of 4,087 patients diagnosed with cancer within a year after their first fatigue record, 47% were diagnosed in the first three months. The number of excess cancer cases among patients with fatigue was greatest in the first month after the index fatigue record, when 856 new cases were observed, compared to 194 expected cases ($p < 0.001$). There followed a steep decrease until month nine, after which the observed monthly cancer cases was similar to expected (month 10 $p = 0.77$) (Figure 4.4, see also Supplementary Appendix 10.4.3 for follow up to 24 months). This was mirrored by a steep initial increase in the cumulative rate of excess cases. By month nine, in patients with fatigue, there were 14 cancer cases per 1,000 patients, compared to an expected 7 per 1,000. By month twelve, there were 16 observed cases per 1,000 patients, compared to 9 expected cases per 1,000 patients (Figure 4.5, Supplementary Appendix 10.4.3).

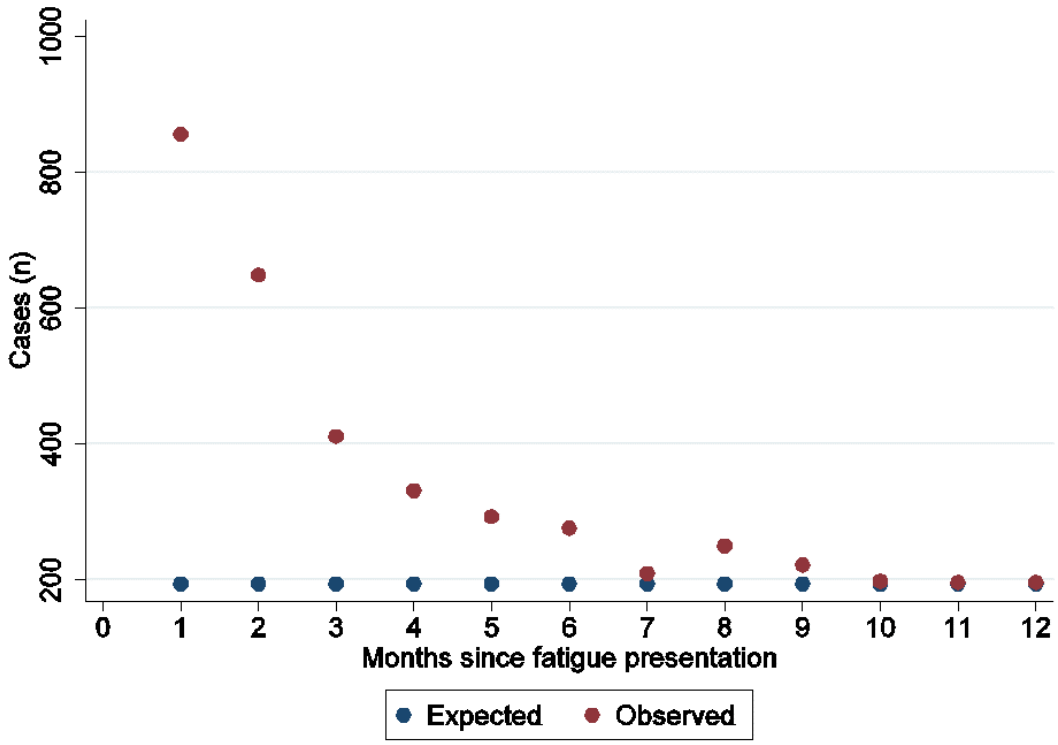



*Figure 4.4. Number of cancer cases by month after first presentation with fatigue, compared to patients in England*

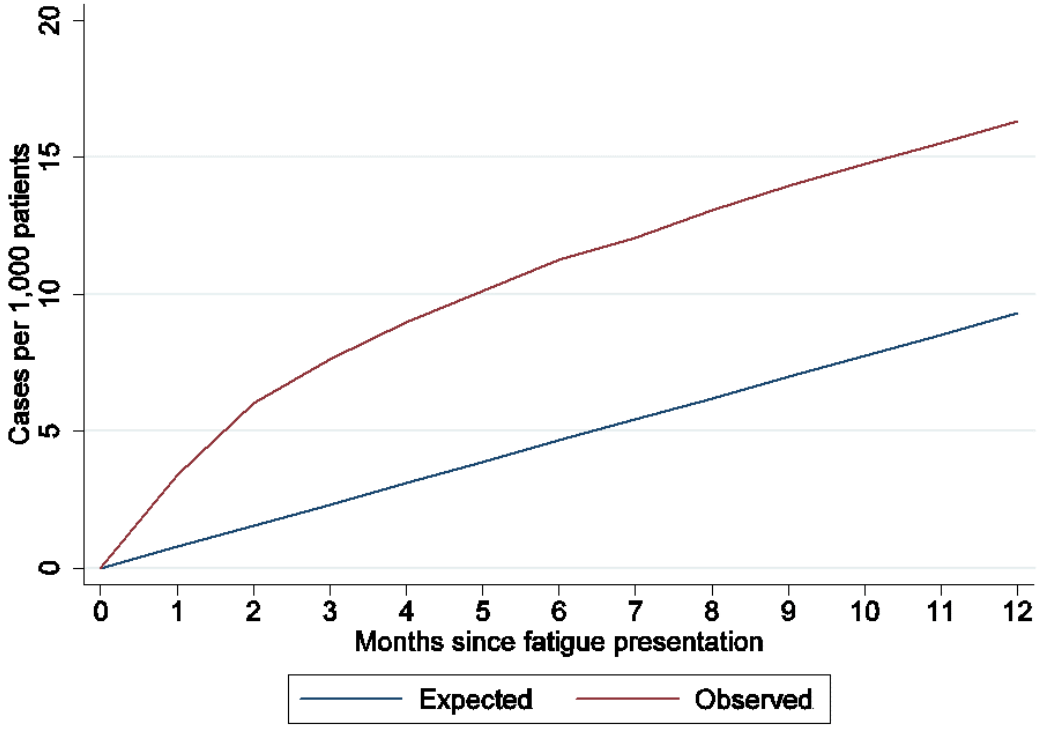



*Figure 4.5. Monthly cumulative rate of cases per 1,000 patients after first presentation with fatigue, compared to patients in England*

### 4.7.5 Sensitivity analyses

Sensitivity analysis showed that cancer risk was similar regardless of whether patients with either a cancer diagnosis or another fatigue record in the previous year were included, or whether the look-back period for exclusion was extended to two years (Table 4.4). In further sensitivity analysis, excluding patients with Chronic Fatigue Syndrome (CFS) or Post Viral Fatigue Syndrome (PVFS) codes from analysis produced similar results overall compared to including them, although cancer risk was lower in patients with CFS or PVF than in other patients with fatigue (Supplementary Appendix 10.4.10).

***Table 4.4. Sensitivity analysis of risk of subsequent cancer diagnosis within 3-24 months after first fatigue presentation, including and excluding patients with previous 'ineligible' fatigue presentations or cancer diagnoses***

*Risk of subsequent cancer diagnosis within 3-24 months after first (index) presentation to primary care with fatigue, including versus excluding eligible fatigue presentations with a previous 'ineligible' fatigue presentation or cancer diagnosis in the previous one or two years.*

| Subsequent cancer | Look-back period: one year[a] | | | | | | | | Look-back period: two years[b] | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Including patients with a previous 'ineligible' fatigue presentation or cancer diagnosis | | Excluding previous 'ineligible' fatigue presentation | | Excluding previous cancer diagnosis | | Excluding previous 'ineligible' fatigue presentation or cancer diagnosis | | Including patients with a previous 'ineligible' fatigue presentation or cancer diagnosis | | Excluding previous 'ineligible' fatigue presentation | | Excluding previous cancer diagnosis | | Excluding previous 'ineligible' fatigue presentation or cancer diagnosis | |
| | n | % | n | % | n | % | n | % | n | % | n | % | n | % | n | % |
| Within 3 months | 2,004 | 0.8% | 1,981 | 0.8% | 1,941 | 0.8% | 1,916 | 0.8% | 1,705 | 0.8% | 1,658 | 0.8% | 1,634 | 0.7% | 1,585 | 0.8% |
| Within 6 months | 2,921 | 1.1% | 2,892 | 1.1% | 2,848 | 1.1% | 2,816 | 1.1% | 2,514 | 1.1% | 2,434 | 1.1% | 2,422 | 1.1% | 2,343 | 1.1% |
| Within 12 months | 4,233 | 1.7% | 4,182 | 1.7% | 4,142 | 1.6% | 4,087 | 1.6% | 3,652 | 1.6% | 3,526 | 1.7% | 3,530 | 1.6% | 3,400 | 1.6% |
| Within 18 months | 5,431 | 2.1% | 5,366 | 2.1% | 5,325 | 2.1% | 5,254 | 2.1% | 4,715 | 2.1% | 4,541 | 2.1% | 4,569 | 2.1% | 4,392 | 2.1% |
| Within 24 months | 6,520 | 2.5% | 6,452 | 2.5% | 6,404 | 2.5% | 6,328 | 2.5% | 5,689 | 2.5% | 5,467 | 2.6% | 5,529 | 2.5% | 5,301 | 2.5% |
| Total patients | 256,865 | | 253,592 | | 254,026 | | 250,606 | | 224,254 | | 213,865 | | 219,947 | | 209,406 | |

[a]Cancer diagnoses between 2007-2015, for patients presenting to primary care with fatigue in 2007-2013. [b]Cancer diagnoses between 2008-2015, for patients presenting to primary care with fatigue in 2008-2013.

## 4.8 Discussion

### 4.8.1 Key findings

The risk of cancer diagnosis within a year following a primary care consultation with fatigue exceeded 3% among men aged 65 and over and women aged 80 and over, and 6% in men aged 80 and over. Cancer risk was at least two-fold greater than that of the general population in men across all age groups, and from 1.5- to 1.7-fold greater than that of the general population in women aged 60 years and over. Although the risk was greater than expected for most cancers, certain cancers, such as leukaemia, pancreatic cancer, and brain cancers, were over-represented among patients with fatigue. Cancer risk was greatest in the three months following the initial presentation, and was three times higher than expected in the first month, but returned to the background rate by nine months.

### 4.8.2 Strengths and limitations

This study used high quality primary care records from CPRD, which is broadly representative of the UK population regarding age, sex, and ethnicity, although may not be representative of all GP practices based on geography and size(145). Full coverage of cancer diagnoses for the study cohort was possible, via linkage to 'gold standard' population-level cancer registration (NCRAS) data (136). The large cohort produced precise estimates of cancer risk by sex and age band, although estimates for rarer cancer sites (e.g. head and neck cancers) may have lacked precision for comparisons between observed and expected risk.

Some instances of a patient's presentation with fatigue may not be recorded by the GP, due to variation in coding practices. GPs are more likely to record alarm symptoms as coded entries rather than free text (which is not generally available to researchers) when there is a suspicion for cancer(89). If coded recording of alarm symptoms is more common in patients who are subsequently diagnosed with cancer than for those who are not, this would artificially inflate cancer risk estimates for alarm symptoms. Under-recording of abdominal pain (which, like fatigue, is a non-specific symptom) was similar in people with and without cancer, which meant that cancer risk estimates were not inflated(89). It is unknown whether similar patterns of under recording apply to fatigue, but if so, reported risk estimates would be representative of the subgroup of fatigue presenters where the GP had greater suspicion of cancer or other serious underlying pathology, and not necessarily representative of the broader group of consultees with fatigue.

My study focuses on patients with fatigue who have sought medical help, and is not generalisable to people experiencing fatigue in the community(26). The comparisons I have made to the general population should be interpreted as contextualising cancer risk among patients presenting to GPs with fatigue, above what would usually be expected for their age and sex. Increased cancer risk may in part reflect differences in the characteristics of patients who consult primary care(26).

Cancer risk in the general population could also be slightly overestimated, as I assumed that one cancer case in the published estimates equalled one person (i.e. that there were no persons with

multiple primary diagnoses). Population incidence estimates also include cancers diagnosed in patients who have presented with fatigue, again making my comparisons of observed versus expected incidence conservative.

To produce risk estimates relevant to primary care clinicians, I aimed to ensure the study population broadly represented patients attending primary care with new-onset fatigue, minimising the likelihood that it was attributable to a previously diagnosed condition or disease (including cancer) or its treatment. Therefore, I excluded patients if all of their potential index fatigue records occurred within a year following another fatigue record or a cancer diagnosis. Nonetheless, a sensitivity analysis showed that results were similar whether including or excluding these patients, and whether extending the look back period from one year before index record to two years.

I did not investigate fatigue in combination with other potential symptoms that could have been reported in the same or an earlier consultation, or related tests or investigations. In common with studies using electronic health records, it is not possible to infer whether the patient's concern about fatigue was the primary reason for the encounter. It is therefore possible that some diagnoses were the result of investigations triggered by another potential cancer sign or symptom in the same consultation or an earlier consultation. This could explain the short time interval between first fatigue presentation and cancer diagnosis in a number of cases. Finally, the date of cancer diagnosis is defined by NCRAS according to hierarchical rules recommended by the European Network of Cancer Registries. In some cases, this is the date of pathological verification, and may be later than the date the patient received the clinical diagnosis of cancer(159).

### 4.8.3 Comparison with literature

In this study, I found that in fatigue presenters, risk was greater than expected for most cancers and it did not provide a strong signal for any particular cancer, although certain cancers, such as leukaemia, pancreatic cancer, and brain cancers, were over-represented among patients with fatigue. Prior literature (largely from case-control studies) has established that fatigue is a known prodromal symptom for many cancers, including lung, colorectal, pancreatic, leukaemia, lymphoma, prostate, renal, ovarian cancers, and a range of childhood cancers (28–36), with proportions of patients presenting with or reporting fatigue ranging from 4% to 45%, depending on the cancer site and study(31). I am not aware of any prior evidence that has established fatigue as a prodromal symptom specifically for brain cancer. This may reflect a gap in the evidence rather than a lack of genuine association.

Available evidence underpinning current NICE guidelines has so far only examined the positive predictive value (PPV) of fatigue for diagnosis of a small number of specific cancer sites(28,29,43), largely because available studies used case-control designs that identified symptoms that were more frequently recorded before diagnosis in patients with a specific cancer, compared to healthy matched controls(29). My study substantially enhances previous evidence regarding the risk of present but as-yet-undetected cancer among patients presenting to primary care with fatigue, as it is the first to examine risk of cancer overall (and by cancer site), establishing that overall cancer risk exceeds referral thresholds in older men and women presenting with fatigue(161)However, a widely-used risk prediction tool (QCancer) reported that fatigue was not a significant independent predictor of cancer within 24-months, unlike other non-site specific symptoms, such as weight loss, appetite loss, and venous thrombo- embolism(44,45). Differences to my study could arise from

various factors, including differences in the data source, length of follow-up, and adjustment for other presenting symptoms.

Few previous studies have sought to identify the most appropriate follow-up period to calculate subsequent cancer risk, though 12 or 24 month periods have been mostly used. One study demonstrated that patients presenting with weight loss (also a non-specific symptom) were at increased risk of a cancer diagnosis up to three months after initial presentation, with rapidly waning risk thereafter(97). My findings mirror this, as half of patients with underlying cancer were diagnosed in the first three months, although observed cancer risk remained substantially higher than expected for patients with fatigue for up to nine months after the index fatigue record.

### 4.8.4 Implications

My study showed that overall one-year cancer risk in patients presenting to primary care with new-onset fatigue was under 3% in men under 65 years, and women under 80. This suggests that, according to current guidelines, urgent two-week-wait referral for suspected cancer would not usually be necessary in these patients if simply considering the presence of fatigue. Notably, cancer risk in younger men (aged 50-64 years) and women (aged 60-75 years) presenting with fatigue was still relatively high compared to the general population. Patients deemed to be at low but not no risk of cancer should still be assessed in primary care, and where necessary investigated for suspected cancer via other urgent or non-urgent pathways, or actively monitored(148). In future, such patients could also become eligible for two-week-wait referral if risk thresholds were to be revised downwards (e.g. to 2%)(160).

Risk was greater than 3% in men aged 65 and over, and women aged 80 and over with new-onset fatigue, suggesting investigation for suspected cancer may be appropriate in these groups. The benefits of ruling out serious physical disease such as cancer must be weighed against the risks of over investigation in older patients with non-specific symptoms, with appropriate communication of diagnostic uncertainty and consideration of patient preferences(161).

In practice, patients with fatigue who also present with a site-specific ‘alarm’ symptom for cancer (e.g. breast lump, rectal bleeding, post-menopausal bleeding) are likely to be referred to an urgent two-week-wait pathway for suspected cancer under National Institute for Health and Care Excellence (NICE) Guidelines, and the diagnostic strategy is considerably clearer in these cases. Therefore, future research is needed to establish the risk of cancer in patients presenting with new-onset fatigue who do not also present with co-occurring ‘alarm’ symptoms. For patients with fatigue who do not present with other, organ-specific, symptoms, it would be helpful to investigate combinations of non-specific symptoms (e.g. fatigue in combination with weight loss), as together these could provide clues as to the level of cancer risk and which cancer sites to suspect. The addition of primary care tests (e.g. commonly used blood tests, chest X-ray, quantitative Faecal Immunochemical Test (qFIT)) could also help to assess the risk of various common cancers. In England, such research could support the development of Rapid Diagnostic Centres (RDC), which aim to expedite diagnosis in patients with non-specific symptoms such as fatigue(46).

Consistent with prior evidence, there were more women than men with fatigue identified in my CPRD population(21,24,44,45,162), which may reflect higher prevalence of conditions (other than cancer) associated with fatigue in women than men(162). Alternatively, help-seeking behaviours

may be different, with men being less likely to report potential cancer symptoms to primary care(26), resulting in an overrepresentation of men with severe fatigue indicating serious underlying physical disease such as cancer. These mechanisms could explain why the observed risk in women was lower than that in men.

The findings relating to the relative frequency of cancer sites diagnosed (i.e. fatigue's 'cancer site signature'(31)) can support the choice of suitable diagnostic test strategies (e.g. the ordering of tests) to most efficiently establish or rule out suspicion of the most likely cancers, when further investigation is deemed appropriate. My study reveals that the case mix of cancers in patients who presented with fatigue is different to that of incident cancer cases in the general population, although the most common cancers still accounted for a large proportion of cases. No cancer site specific risk exceeded the NICE 3% two-week-wait referral threshold, although leukaemia, pancreatic and brain cancers were particularly overrepresented among patients with fatigue, relative to their expected incidence. This could reflect cancer-specific pathophysiological mechanisms, for example, a high prevalence of anaemia leading to fatigue as a presenting symptom in patients with leukaemia. However, I could not examine such biological pathways directly, and other explanatory mechanisms may be possible.

The findings suggest that should a clinician and patient decide to 'actively monitor' any potential cancer risk following the patient's first presentation with fatigue, the length of this period should be up to nine months, though most of this risk is concentrated in the first three. These represent periods when both healthcare professionals and patients should be vigilant of developing symptoms for cancer – though it should also be borne in mind that other (non-neoplastic diseases) may also develop. Future research could establish whether this period of excess risk varies by cancer site. For the main analyses, I provided one-year cancer risk estimates to facilitate comparison with existing NICE Guidelines. As the majority of excess cases occur soon after fatigue presentation, the difference between nine and twelve-month risk estimates was small (0.2 percentage points, overall).

## 4.9 Chapter summary

In this chapter, I established that in men over 65 and women over 80, presenting to primary care with fatigue was associated with cancer risk that exceeds current thresholds for urgent two-week wait investigation. This demonstrated that fatigue warranted further exploration as an indicator of undetected cancer, with subsequent chapters needed to consider how risk is modified by the presence or absence of other signs and symptoms. Fatigue was associated with a broad range of cancer sites, but is not strongly predictive of any specific one, though certain cancers were more likely. Cancer risk was elevated among patients with fatigue for up to nine months after initial presentation, establishing an appropriate follow up period for subsequent chapters.

# 5. Chapter 5: Underlying cancer risk among patients with fatigue and other vague symptoms in primary care: a population-based cohort study

## 5.1 Chapter rationale

While it is clear that patients presenting with fatigue who also present with an 'alarm' symptom for cancer (e.g. breast lump, rectal bleeding) should be urgently referred for suspected cancer, the referral strategy is considerably less clear for patients with fatigue either as the sole presenting symptom or in combination with potential cancer features that, like fatigue, are non-alarm or non-site specific (e.g. weight loss, abdominal pain, anaemia). In this chapter, I therefore aimed to estimate the risk of incident diagnosis of any cancer in patients who present with new-onset fatigue without accompanying alarm symptoms for cancer, according to combinations of other presenting vague symptoms.

## 5.2 Publication

This chapter has been published in the peer reviewed journal, British Journal of General Practice:

White, B., Renzi, C., Barclay, M., & Lyratzopoulos, G. (2023). Underlying cancer risk among patients with fatigue and other vague symptoms: a population-based cohort study in primary care. British Journal of General Practice, BJGP.2022.0371. https://doi.org/10.3399/BJGP.2022.0371

## 5.3 Author contributions

Authors: Becky White, Cristina Renzi, Matthew Barclay, Georgios Lyratzopoulos

BW, GL, and CR conceived and designed the study. BW managed and analysed the data, with statistical analyses and graphical presentation supervised/ developed by MB. CR and GL provided clinical input. All authors contributed to drafting and revising the article. Symptoms were defined using libraries of Read codes developed by Prof Willie Hamilton (WH) and Dr Sarah Price (SP) at Exeter University, with additional codes added by colleagues GL, CR, BW, MB, and Dr Meena Rafiq (MR) at UCL.

 For more information, including author contributions, see Appendix 10.5.1.

## 5.4 Abstract

**Background**

Presenting to primary care with fatigue is associated with slightly increased cancer risk, although it is unknown how this varies in the presence of other 'vague' symptoms.

**Aim**

To quantify cancer risk in fatigued patients presenting with other 'vague' symptoms, in the absence of 'alarm' symptoms for cancer.

**Design and Setting**

Cohort study of patients presenting in UK primary care with new-onset fatigue during 2007-2015, using Clinical Practice Research Datalink data linked to national cancer registration data.

**Method**

I identified fatigue presenters without co-occurring alarm symptoms or anaemia, whom I further characterised for co-occurrence of 19 other 'vague' potential cancer symptoms. I calculated sex and age-specific nine-month cancer risk for each fatigue-vague symptom cohort.

**Results**

Of 285,382 patients presenting with new-onset fatigue, 84% (n=239,846) did not have co-occurring alarm symptoms or anaemia. Of these, 38% (n=90,828) presented with at least one of 19 vague symptoms for cancer.

Cancer risk exceeded 3% in older men with fatigue combined with any of the vague symptoms studied. The age at which risk exceeded 3% was 59 years for fatigue-weight loss, 65 years for fatigue-abdominal pain, 67 years for fatigue-constipation, and 67 years for fatigue-other upper gastro-intestinal symptoms. For women, risk exceeded 3% only in older patients with fatigue-weight loss (from 65 years), fatigue-abdominal pain (from 79 years), or fatigue-abdominal bloating (from 80 years).

**Conclusion**

In the absence of alarm symptoms or anaemia, fatigue combined with specific vague presenting symptoms, alongside patient age and sex, can guide clinical decisions about referral for suspected cancer.

## 5.5 Background

Many cancer patients are diagnosed after presenting with vague symptoms(17), such as fatigue, which are characterised by lack of organ-specificity and low positive predictive value (PPV) for any single cancer type. Vague symptoms are not generally supported by urgent referral recommendations for suspected cancer under UK National Institute for Health and Care Excellence (NICE) Guidelines, except for some specific patient groups and cancer sites. Patients diagnosed with cancer following presentation with these symptoms typically experience prolonged diagnostic intervals(13).

Fatigue is a relatively common presenting symptom in primary care, being the primary complaint in an estimated 5-7% of consultations(21–24), and more commonly reported by women than men(21,24,162). It presents a diagnostic challenge, particularly regarding assessing the risk of underlying cancer(22,23,39,40,42). Although fatigue is reported by patients before diagnosis for a number of cancer sites(28–36), its predictive value for any single cancer site is low(29,38). Fatigue could also signal many other conditions, including self-limiting illnesses (e.g. short-term post-viral fatigue), depression, chronic fatigue syndrome, autoimmune disease (e.g. lupus), chronic infection (e.g. hepatitis C), or a range of other causes (e.g. hypothyroidism, vitamin deficiency, iron deficiency, coeliac disease etc.)(22,39–42).

When new-onset fatigue accompanies an 'alarm' symptom for cancer, diagnostic management is typically straightforward. For example, in England, patients with 'alarm' symptoms for cancer can be referred to appropriate hospital specialties for urgent ('two-week-wait') investigation for suspected cancer (as per guidelines published by NICE)(28,29,43). However, when patients with new-onset fatigue present with vague symptoms only, diagnostic management is less clear. For the purposes of this thesis, I refer to potential cancer symptoms that are not listed in NICE Guidelines for suspected cancer as 'vague' symptoms; all of these symptoms likely have a low predictive value for any single cancer site, and most are associated with a broad range of cancer sites, but it should be noted that some (e.g. UTIs) are more organ-specific than others.

When patients present with fatigue with vague symptoms only, GPs must discern which of these patients should nevertheless be investigated for cancer due to elevated risk associated with their demographic group or other vague signs and symptoms combined with fatigue, and whether to refer onto an urgent ('two-week-wait') pathway for suspected cancer, or to a multidisciplinary diagnostic centre ('Rapid Diagnostic Centres' (RDCs) in England).

More detailed evidence is needed to support such decision-making. In a previous study I quantified the risk of cancer diagnosis shortly after new-onset fatigue(163). However, how often fatigue presents alongside other symptoms, and the associated risk of underlying cancer, is not known, although similar studies have been conducted in cohorts of patients with other vague symptoms, including weight loss or abdominal symptoms(48,49,164). Current evidence assessing cancer risk in patients with fatigue in combination with other presenting features is limited to specific cancer sites(30–32,35,38,51) or symptom combinations(49,52). Furthermore, a detailed examination of cancer risk in patients presenting with new-onset fatigue in the absence of alarm symptoms would support GPs to identify which patients to refer in a group of patients for whom diagnostic management is particularly challenging.

Therefore, I aimed to estimate the short-term risk of incident diagnosis of any malignant neoplasm (excluding non-melanoma skin cancer) in patients who present with new-onset fatigue without

accompanying alarm symptoms for cancer, according to combinations of other presenting vague symptoms.

## 5.6 Methods

### 5.6.1 Study design and data source

I conducted a cohort study of patients with a record of fatigue presentation in primary care in England between January 2007 and April 2015, using electronic health records (EHRs) from the Clinical Practice Research Datalink (CPRD) GOLD (March 2019 database build), linked to Index of Multiple Deprivation (IMD) quintile, and cancers diagnosed from 2006-2015 in National Cancer Registration and Analysis Service (NCRAS) data. For more information about CPRD and linked datasets, see Section 3.6.2.

### 5.6.2 Symptom identification

In addition to fatigue, I identified 64 'potential' cancer symptoms from those listed in NICE 2015 diagnostic guidelines for suspected cancer(28,29,43) and additional sources(46,165,166). Additional symptoms of interest did not need to be established by prior literature as fatigue-related. Read code lists were available for 35 of the identified symptoms, which were therefore included in the study(143) (12,13,79,94,143,167–172).

Of the additional 35 symptoms included in the study, 16 were categorised as 'alarm', defined as those with NICE NG 12 (2015) recommendations for urgent two-week wait referral or investigation for suspected cancer(28,29,43). The remaining 19 symptoms were categorised as 'vague' (Figure 5.1). Appendix 10.5.2 lists the sources used to define each symptom, including fatigue, with all Read codes available at https://github.com/rmjlrwh/Fatigue. Of the 28 potential cancer symptoms that were not profiled due to unavailable Read code lists, 12 were categorised as 'alarm' and 16 as 'vague'. These are listed in Appendix 10.5.3.

Fatigue presenters without an alarm symptom but with anaemia (defined as a low haemoglobin test result, using published methods(13,173) (Appendix 10.5.4)) were analysed separately, as anaemia in older patients would usually prompt urgent referral under NICE 2015 diagnostic guidelines(28,29,43).

### 5.6.3 Cohort identification

First, a cohort of patients aged 30-99 years presenting to primary care with new-onset fatigue between 2007-2015, and no cancer diagnosis in the previous year was identified in CPRD (Figure 5.1). The steps taken to define this cohort are detailed in a previous publication(163).Section 4.6.2, with more detail about the rationale detailed in Section 3.6.5.

Patients with fatigue with a 'co-occurring' alarm symptom (occurring between three months prior to one month after the first fatigue presentation) were excluded from subsequent age-specific analysis. Patients with fatigue and no alarm symptoms were characterised for presence of 'co-occurring'

anaemia. Finally, for patients with fatigue and no alarm symptoms or anaemia, subcohorts of patients with fatigue and each co-occurring vague symptom were identified. These cohorts were not mutually exclusive, i.e. the same patient could be in more than one cohort if they had more than one symptom combined with fatigue.

A time window of three months prior to one month after the first fatigue presentation was chosen to define 'co-occurrence', because patients' diagnostic episodes could span multiple visits to the doctor over a short period of time, and doctors may not record all presenting symptoms during each consultation. Records of additional symptoms or anaemia were considered 'eligible' if meeting criteria detailed in Figure 5.1.

### 5.6.4 Follow up and outcomes

Follow up began with the patient's first eligible record of fatigue during the study period, and ended either at nine months, or the first cancer diagnosis, if earlier. As NCRAS data was used to define the outcome, patients could remain in the study even after they left their GP practice or their practice exited CPRD. Patients could not subsequently re-enter the study with another fatigue record.

The main outcome was diagnosis of cancer recorded in NCRAS data within nine months following the first fatigue presentation. Nine months was chosen following a previous publication's findings that excess cancer risk is concentrated in this period(163). Cancers included any malignant neoplasms, excluding non-melanoma skin cancer (International Classification of Diseases 10th Revision (ICD-10) codes C00-C99 excl. C45). Benign brain tumours were not included(163).

### 5.6.5 Statistical analyses

I calculated cancer risk for patients with and without alarm symptoms or anaemia, and in the cohort of patients without, I calculated risk for each 'fatigue-co-occurring vague symptom' subcohort. Analysis was stratified by sex, but not age band, due to sample size constraints. For instance, in the 'fatigue and weight loss' subcohort, there were under 50 men in each five year age band, and no cancer cases below the age of 60, so risk estimates in these age bands would be very imprecise. Instead, I fitted Poisson regression models, with cancer diagnosis as the outcome and age modelled as a continuous exposure variable using restricted cubic splines, and produced modelled cancer risk at selected ages. Robust standard errors were used to account for possible overdispersion. I plotted residuals to ascertain model fit in each co-occurring symptom group. Potential interactions were observed between age and weight loss, and age and abdominal bloating (women only), but the addition of interaction terms did not improve model fit, so these were not included. Due to small sample sizes, pelvic pain and night sweats were not included in age-specific analyses.

To contextualise modelled age-specific cancer risk estimates, I also showed nine-month cancer risk in the general population (derived using incident cancer registration statistics for England in

2011(151) and corresponding mid-year population estimates)(152). Due to data availability, these were for five-year age bands, and all ages from 85 years were grouped together.

Data management and analysis was conducted in MySQL Workbench v6.1 and Stata v17, respectively. All relevant code is available online at https://github.com/rmjlrwh/Fatigue. I used the Strengthening the Reporting of Observational studies in Epidemiology (STROBE) guidelines for cohort studies(158).

### 5.6.6 Sensitivity analyses

A sensitivity analysis examined the impact on cancer risk estimates of varying the time window used to define symptom co-occurrence before the first fatigue presentation, up to 12 months pre-presentation.

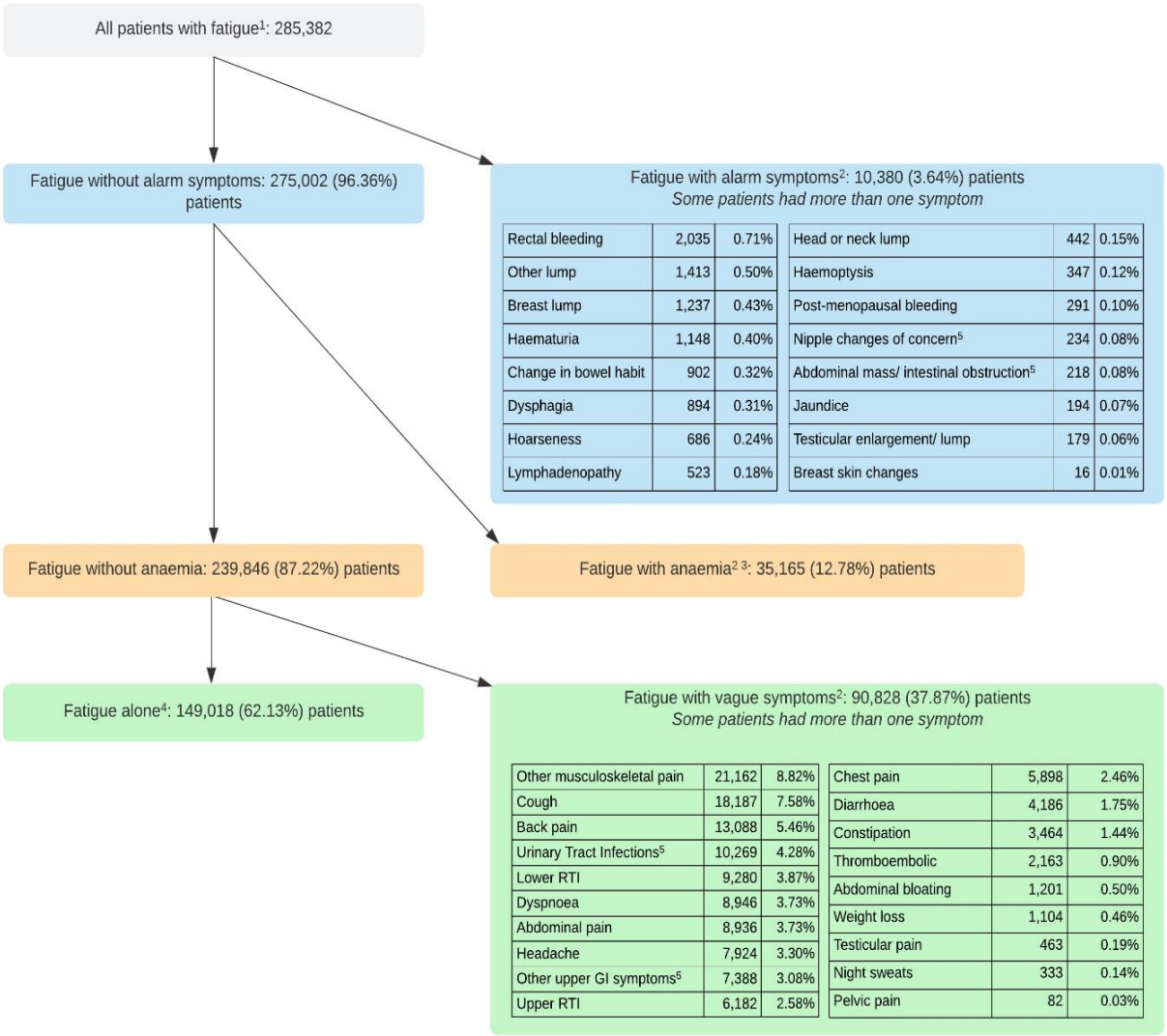
## 5.7 Findings

### 5.7.1 Cohort inclusions and exclusions

285,382 patients had at least one 'eligible' record of fatigue in primary care within the patient's inclusion period, without a cancer diagnosis or fatigue record in the previous year (Figure 5.1).

10,380 (3.6%) patients with fatigue had a co-occurring alarm symptom three months before to one month after their first eligible fatigue record. Of the remaining patients, 35,165 (12.8%) had anaemia.

Overall, 239,846 (84%) patients with fatigue did not have any alarm symptoms or anaemia. Of these (N=239,846), 90,828 (38%) had one or more co-occurring vague symptoms. Approximately half (52%, n=149,018) of all patients with fatigue had fatigue alone i.e. all other potential (alarm and vague) cancer symptoms studied were absent.




| | | | | | |
|---|---|---|---|---|---|
| Rectal bleeding | 2,035 | 0.71% | Head or neck lump | 442 | 0.15% |
| Other lump | 1,413 | 0.50% | Haemoptysis | 347 | 0.12% |
| Breast lump | 1,237 | 0.43% | Post-menopausal bleeding | 291 | 0.10% |
| Haematuria | 1,148 | 0.40% | Nipple changes of concern[5] | 234 | 0.08% |
| Change in bowel habit | 902 | 0.32% | Abdominal mass/ intestinal obstruction[5] | 218 | 0.08% |
| Dysphagia | 894 | 0.31% | Jaundice | 194 | 0.07% |
| Hoarseness | 686 | 0.24% | Testicular enlargement/ lump | 179 | 0.06% |
| Lymphadenopathy | 523 | 0.18% | Breast skin changes | 16 | 0.01% |

| | | | | | |
|---|---|---|---|---|---|
| Other musculoskeletal pain | 21,162 | 8.82% | Chest pain | 5,898 | 2.46% |
| Cough | 18,187 | 7.58% | Diarrhoea | 4,186 | 1.75% |
| Back pain | 13,088 | 5.46% | Constipation | 3,464 | 1.44% |
| Urinary Tract Infections[5] | 10,269 | 4.28% | Thromboembolic | 2,163 | 0.90% |
| Lower RTI | 9,280 | 3.87% | Abdominal bloating | 1,201 | 0.50% |
| Dyspnoea | 8,946 | 3.73% | Weight loss | 1,104 | 0.46% |
| Abdominal pain | 8,936 | 3.73% | Testicular pain | 463 | 0.19% |
| Headache | 7,924 | 3.30% | Night sweats | 333 | 0.14% |
| Other upper GI symptoms[5] | 7,388 | 3.08% | Pelvic pain | 82 | 0.03% |
| Upper RTI | 6,182 | 2.58% | | | |

***Figure 5.1. Study cohorts***

*1 Patients with at least one eligible fatigue record in CPRD between 01/01/2007 - 02/04/2015. Fatigue records were eligible if occurring after the practice was 'up to standard' and the patient was registered to the practice for > 1 year, the patient was 30 years+, before the practice's last collection date, the patient left the practice, turned 100 years, or died. There also had to be no fatigue record or cancer diagnosis within the previous year.*
*2 Symptoms/ tests were 'co-occurring' i.e. recorded 3 months before - 1 month after the patient's first eligible fatigue record. Co-occurring symptoms/ tests were eligible if occurring after the practice was 'up to standard' and the patient was registered to the practice, and before the practice's last collection date, the patient left the practice, died, or was diagnosed with cancer.*
*3 Patients had at least one valid low haemoglobin measurement meeting the above eligibility criteria, and the measurement was considered valid (i.e. within a biologically plausible range)*
*4 Without any of the studied alarm or vague symptoms, or anaemia.*
*5 Abdominal mass/ intestinal obstruction also includes rectal mass. Nipple changes of concern also include nipple discharge or retraction. Urinary Tract Infection also includes cystitis, dysuria, urgency, painful urination, urine smell. Other upper GI (gastro-intestinal) symptoms include dyspepsia, nausea, vomiting, haematemesis, loss of appetite.*

### 5.7.2 Frequency of co-occurring vague symptoms

Among patients with no alarm symptoms or anaemia (N=239,846), the five most common vague symptom combinations were fatigue-musculoskeletal pain, fatigue-cough, fatigue-back pain, fatigue-dyspnoea, and fatigue-lower respiratory tract infections (Figure 5.2). Of patients with fatigue and no alarm symptoms or anaemia (N=239,846), 26% (n=62,732) had only one additional type of vague symptom in combination with fatigue, and 12% (n=28,096) had two or more (e.g. fatigue with abdominal pain and cough) (Appendix 10.5.5). The cohort size and median age (IQR) of the studied vague symptom combinations with fatigue are presented in Table 5.1.

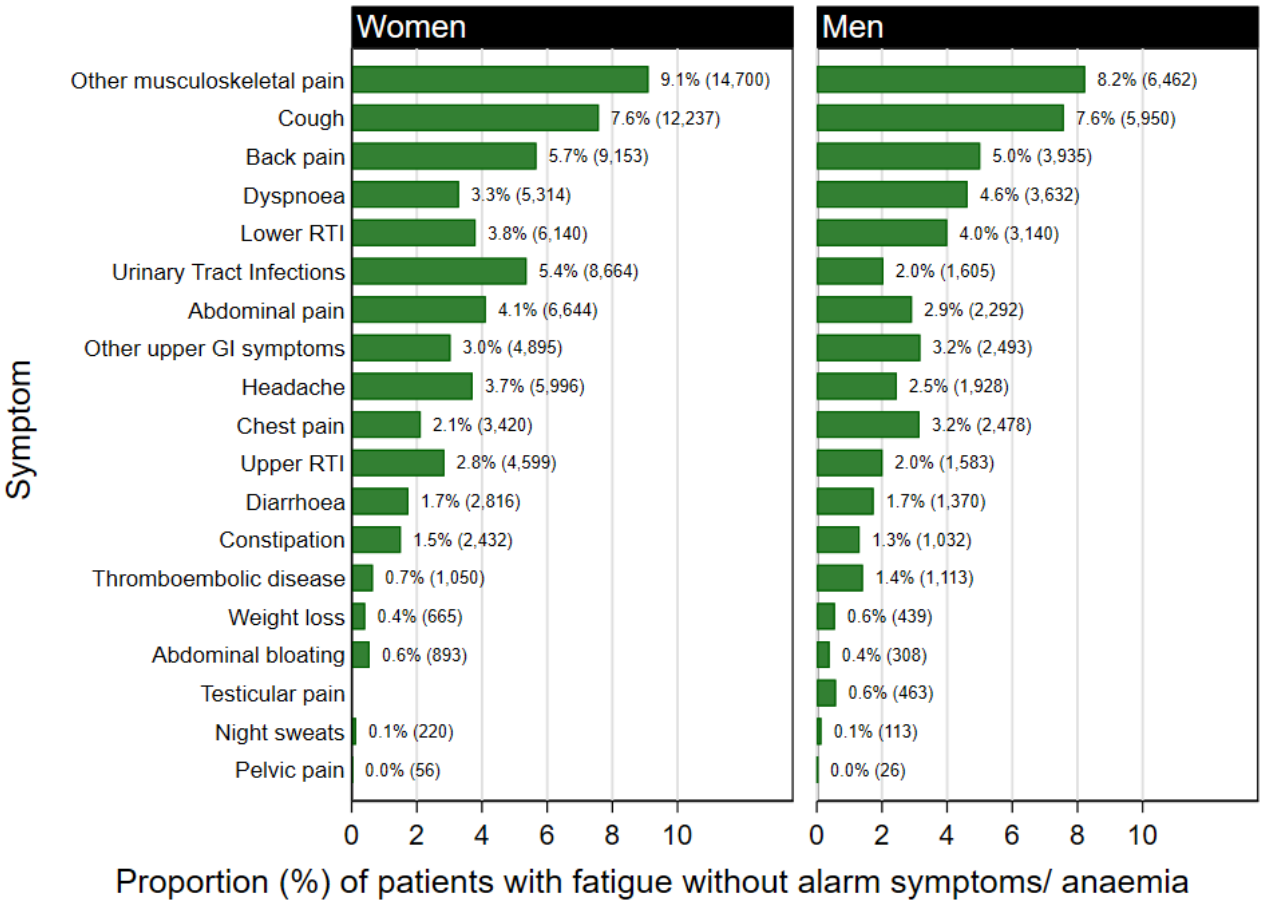



***Figure 5.2. Patients with each co-occurring* vague symptom, as a proportion of patients with fatigue and no alarm symptoms or anaemia (%)***

**Co-occurring symptoms were those recorded 3 months before – 1 month after the first fatigue presentation. These cohorts were not mutually exclusive; 12% of patients had more than one of these vague symptoms. Urinary Tract Infections also include cystitis, dysuria, urgency, painful urination, urine smell. Other upper GI (gastro-intestinal) symptom includes dyspepsia, nausea, vomiting, haematemesis, loss of appetite. RTI = Respiratory Tract Infections.*

***Table 5.1. Age characteristics of patients with fatigue, with each co-occurring* symptom***

*For a) all patients with fatigue b) patients with fatigue, without alarm symptoms, and c) patients with fatigue, without alarm symptoms or anaemia*

| | Women | | Men | |
|---|---|---|---|---|
| | Total (N) | Median (IQR) age (years) | Total (N) | Median (IQR) age (years) |
| **a) All patients with fatigue** | **192,614** | **52 (41-69)** | **92,768** | **58 (46-71)** |
| With alarm symptoms | 6,916 | 53 (42-69) | 3,464 | 63 (48-76) |
| **b) Patients with fatigue, without alarm symptoms** | **185,698** | **52 (41-69)** | **89,304** | **58 (46-71)** |
| With anaemia | 24,323 | 59 (43-78) | 10,833 | 76 (66-83) |
| **c) Patients with fatigue, without alarm symptoms or anaemia** | **161,375** | **52 (41-67)** | **78,471** | **56 (44-68)** |
| With vague symptoms | 62,300 | 56 (43-71) | 28,528 | 59 (47-72) |
| Without vague symptoms (i.e. fatigue only) | 99,075 | 50 (40-64) | 49,943 | 54 (43-65) |
| *Pairwise combinations of fatigue with each vague symptom:* | | | | |
| Abdominal pain | 6,644 | 51 (40-66) | 2,292 | 57 (45-69) |
| Abdominal bloating | 893 | 53 (42-69) | 308 | 59 (46-70) |
| Dyspnoea | 5,314 | 68 (55-78) | 3,632 | 68 (57-77) |
| Night sweats | 220 | 53 (44-65) | 113 | 57 (49-67) |
| Weight loss | 665 | 65 (48-79) | 439 | 63 (50-76) |
| Constipation | 2,432 | 65 (46-80) | 1,032 | 71 (60-80) |
| Cough | 12,237 | 58 (45-71) | 5,950 | 61 (48-72) |
| Diarrhoea | 2,816 | 60 (44-76) | 1,370 | 59 (46-72) |
| Pelvic pain | 56 | 42 (38-55) | 26 | 55 (43-65) |
| Other Upper GI symptoms | 4,895 | 59 (45-72) | 2,493 | 58 (45-70) |
| Urinary Tract Infections | 8,664 | 60 (44-76) | 1,605 | 70 (55-80) |
| Other musculoskeletal pain | 14,700 | 57 (45-71) | 6,462 | 59 (48-70) |
| Chest pain | 3,420 | 58 (46-72) | 2,478 | 59 (48-71) |
| Testicular pain | | | 463 | 52 (42-65) |
| Headache | 5,996 | 47 (38-59) | 1,928 | 51 (41-62) |
| Back pain | 9,153 | 53 (42-68) | 3,935 | 56 (45-68) |
| Upper RTI | 4,599 | 50 (40-63) | 1,583 | 55 (43-66) |
| Lower RTI | 6,140 | 61 (48-75) | 3,140 | 64 (51-76) |
| Thromboembolic disease | 1,050 | 74 (62-83) | 1,113 | 69 (59-77) |

**Co-occurring symptoms were those recorded 3 months before – 1 month after the first fatigue presentation. Urinary Tract Infections also include cystitis, dysuria, urgency, painful urination, urine smell. Other upper GI (gastro-intestinal) symptom includes dyspepsia, nausea, vomiting, haematemesis, loss of appetite. RTI = Respiratory Tract Infections.*

### 5.7.3 Cancer risk in patients with and without alarm symptoms

For all patients with fatigue (all ages combined), observed cancer risk within nine months after first fatigue presentation was 2.2% [2.1%-2.3%] in men and 1.1% [1.0%-1.1%] in women. Risk was higher for those with alarm symptoms than those without (Appendix **Error! Reference source not found.**, REF _Ref136937509 \r \h * MERGEFORMAT 10.5.6, 10.5.7).

### 5.7.4 Cancer risk in patients with and without anaemia

For patients with fatigue and no alarm symptoms, observed cancer risk was higher for those with anaemia than those without (Appendix **Error! Reference source not found.**, 10.5.6, 10.5.7). M odelled age-specific risk for patients with anaemia exceeded 3% in men from 57 years (3.1% [2.7%-3.6%]) and women from 62 years (3.0% [2.7%-3.4%]), and 8% in men from 71 years (8.1% [7.4%-8.9%]) (Figure 5.3, Figure 5.4, Appendix 10.5.8).

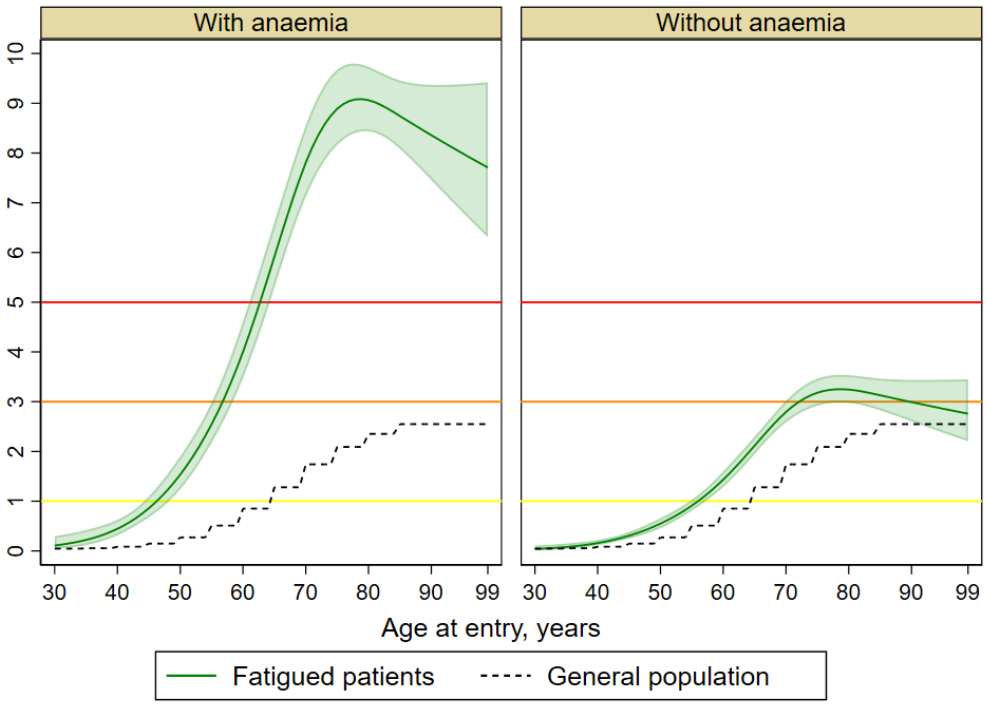



***Figure 5.3. Modelled nine-month cancer risk (%) in male patients with fatigue and no alarm symptoms, for each year of age (30-99 years), by presence of anaemia***

*Risk for non-linear continuous age modelled using restricted cubic splines. Includes observed nine-month cancer risk (%) for the general population in England in 2011, by five year age band. Available population estimates grouped all men aged 85+*

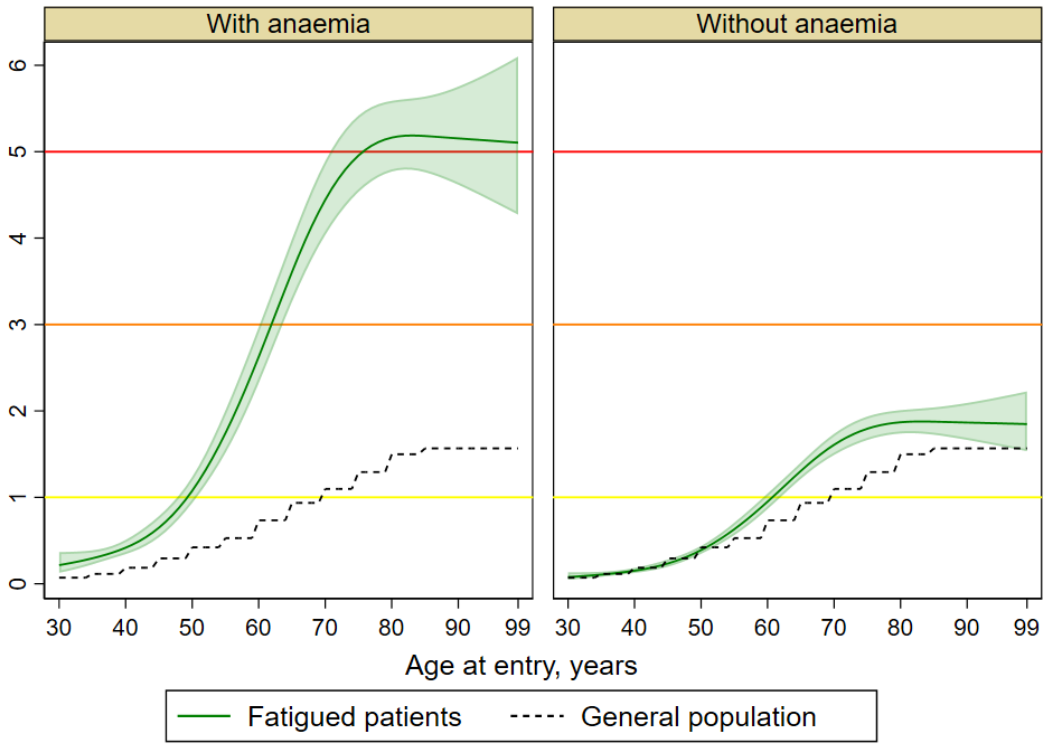



***Figure 5.4. Modelled nine-month cancer risk (%) in female patients with fatigue and no alarm symptoms, for each year of age (30-99 years), by presence of anaemia***

*Risk for non-linear continuous age modelled using restricted cubic splines. Includes observed nine-month cancer risk (%) for the general population in England in 2011, by five year age band. Available population estimates grouped all women aged 85+*

### 5.7.5 Cancer risk in patients with each vague symptom

For patients with fatigue and no alarm symptom or anaemia, observed cancer risk for all ages combined was higher for fatigue presenters with at least one co-occurring vague symptom compared to those without. Cancer risk was higher for patients with two or more different additional vague symptoms in combination with fatigue (men: 2.5% [2.2%-2.9%], women: 1.3% [1.2%-1.5%] for women) compared to those with only one additional vague symptom (men: 1.5% [1.4%-1.7%], women: 0.8% [0.8%-0.9%] for women) (Appendix 10.5.5). For 16 out of 17 fatigue-co-occurring symptom combinations studied in women, and 15 out of 18 in men, at least a third of cancers diagnosed were for cancer sites other than the three most common in that symptom cohort (Appendix 10.5.10).

Overall, for all ages combined, observed cancer risk was highest for weight loss, constipation, dyspnoea, abdominal pain (men), or abdominal bloating (women) (Table 5.2, Appendix 10.5.6). Age-specific modelled cancer risk increased with age for each vague symptom (Figure 5.5, Figure 5.6, Figure 5.7, Figure 5.8). Adjusting for age, cancer risk was higher for fatigue in combination with any vague symptom, compared to fatigue without co-occurring vague symptoms (Appendix 10.5.7). These combinations included four specific symptoms in men, and six in women: fatigue-weight loss,

fatigue-abdominal pain, fatigue-constipation, fatigue-other upper gastro-intestinal (GI) symptoms, fatigue-abdominal bloating (women), or fatigue-dyspnoea (women).

***Table 5.2. Observed cancer risk by each co-occurring symptom***

*Observed nine-month cancer risk (%) for patients with fatigue aged 30-99 years who had a co-occurring symptom 3 months before to 1 month after the first fatigue presentation, for a) all patients with fatigue, b) patients with fatigue without alarm symptoms, c) patients with fatigue without alarm symptoms or anaemia. Cell counts under 5 are suppressed to reduce statistical disclosure risk. Urinary Tract Infections also include cystitis, dysuria, urgency, painful urination, urine smell. Other upper GI (gastro-intestinal) symptoms include dyspepsia, nausea, vomiting, haematemesis, loss of appetite. RTI = Respiratory Tract Infections.*

| | Women | | | | Men | | | |
|---|---|---|---|---|---|---|---|---|
| | Total patients (N) Total | Patients with cancer (n) Total | % | (95% CI) | Total patients (N) Total | Patients with cancer (n) Total | % | (95% CI) |
| **a) All patients with fatigue** | **192,614** | **2,101** | **1.09** | **(1.05, 1.14)** | **92,768** | **2,036** | **2.19** | **(2.10, 2.29)** |
| With alarm symptoms (excluding anaemia) | 6,916 | 208 | 3.01 | (2.63, 3.44) | 3,464 | 194 | 5.60 | (4.88, 6.42) |
| Breast lump | 1,186 | 50 | 4.22 | (3.21, 5.52) | 51 | < 5 | - | - |
| Change in bowel habit | 564 | 24 | 4.26 | (2.88, 6.25) | 338 | 18 | 5.33 | (3.39, 8.26) |
| Nipple changes of concern | 212 | < 5 | - | - | 22 | < 5 | - | - |
| Dysphagia | 527 | 18 | 3.42 | (2.17, 5.33) | 367 | 22 | 5.99 | (3.99, 8.91) |
| Haematuria | 576 | 14 | 2.43 | (1.45, 4.04) | 572 | 35 | 6.12 | (4.43, 8.39) |
| Haemoptysis | 167 | 7 | 4.19 | (2.05, 8.40) | 180 | 15 | 8.33 | (5.11, 13.29) |
| Jaundice | 83 | 22 | 26.51 | (18.20, 36.89) | 111 | 27 | 24.32 | (17.29, 33.08) |
| Post-menopausal bleeding | 291 | 10 | 3.44 | (1.88, 6.21) | | | | |
| Rectal bleeding | 1,193 | 26 | 2.18 | (1.49, 3.17) | 842 | 24 | 2.85 | (1.92, 4.21) |
| Abdominal mass/ intestinal obstruction | 145 | 17 | 11.72 | (7.45, 17.97) | 73 | 14 | 19.18 | (11.78, 29.66) |
| Breast skin changes | 16 | < 5 | - | - | | | | |
| Lymphadenopathy | 382 | 10 | 2.62 | (1.43, 4.75) | 141 | 10 | 7.09 | (3.90, 12.56) |
| Hoarseness | 484 | 7 | 1.45 | (0.70, 2.95) | 202 | 11 | 5.45 | (3.07, 9.49) |
| Head or neck lump | 331 | < 5 | - | - | 111 | 12 | 10.81 | (6.29, 17.95) |
| Testicular enlargement/ lump | | | | | 179 | 6 | 3.35 | (1.55, 7.12) |
| Other lump | 1,016 | 11 | 1.08 | (0.61, 1.93) | 397 | 11 | 2.77 | (1.55, 4.89) |
| **b) Patients with fatigue without alarm symptoms** | **185,698** | **1,893** | **1.02** | **(0.97, 1.07)** | **89,304** | **1,842** | **2.06** | **(1.97, 2.16)** |
| With anaemia | 24,323 | 661 | 2.72 | (2.52, 2.93) | 10,833 | 778 | 7.18 | (6.71, 7.68) |
| **c) Patients with fatigue without alarm symptoms or anaemia** | **161,375** | **1,232** | **0.76** | **(0.72, 0.81)** | **78,471** | **1,064** | **1.36** | **(1.28, 1.44)** |
| With vague symptoms | 62,300 | 619 | 0.99 | (0.92, 1.07) | 28,528 | 519 | 1.82 | (1.67, 1.98) |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Without vague symptoms | 99,075 | 613 | 0.62 | (0.57, 0.67) | 49,943 | 545 | 1.09 | (1.00, 1.19) |
| *Pairwise combinations of fatigue with each vague symptom:* | | | | | | | | |
| Abdominal pain | 6,644 | 96 | 1.44 | (1.18, 1.76) | 2,292 | 61 | 2.66 | (2.08, 3.40) |
| Abdominal bloating | 893 | 16 | 1.79 | (1.11, 2.89) | 308 | 7 | 2.27 | (1.11, 4.62) |
| Dyspnoea | 5,314 | 94 | 1.77 | (1.45, 2.16) | 3,632 | 99 | 2.73 | (2.24, 3.31) |
| Night sweats | 220 | < 5 | - | - | 113 | < 5 | - | - |
| Weight loss | 665 | 22 | 3.31 | (2.19, 4.96) | 439 | 27 | 6.15 | (4.26, 8.80) |
| Constipation | 2,432 | 48 | 1.97 | (1.49, 2.61) | 1,032 | 41 | 3.97 | (2.94, 5.35) |
| Cough | 12,237 | 118 | 0.96 | (0.81, 1.15) | 5,950 | 116 | 1.95 | (1.63, 2.33) |
| Diarrhoea | 2,816 | 26 | 0.92 | (0.63, 1.35) | 1,370 | 32 | 2.34 | (1.66, 3.28) |
| Pelvic pain | 56 | < 5 | - | - | 26 | < 5 | - | - |
| Other upper GI symptoms | 4,895 | 79 | 1.61 | (1.30, 2.01) | 2,493 | 62 | 2.49 | (1.94, 3.18) |
| Urinary Tract Infections | 8,664 | 104 | 1.20 | (0.99, 1.45) | 1,605 | 41 | 2.55 | (1.89, 3.45) |
| Other musculoskeletal pain | 14,700 | 116 | 0.79 | (0.66, 0.95) | 6,462 | 99 | 1.53 | (1.26, 1.86) |
| Chest pain | 3,420 | 45 | 1.32 | (0.98, 1.76) | 2,478 | 52 | 2.10 | (1.60, 2.74) |
| Testicular pain | | | | | 463 | 7 | 1.51 | (0.73, 3.09) |
| Headache | 5,996 | 38 | 0.63 | (0.46, 0.87) | 1,928 | 23 | 1.19 | (0.80, 1.78) |
| Back pain | 9,153 | 84 | 0.92 | (0.74, 1.13) | 3,935 | 55 | 1.40 | (1.08, 1.81) |
| Upper RTI | 4,599 | 24 | 0.52 | (0.35, 0.78) | 1,583 | 23 | 1.45 | (0.97, 2.17) |
| Lower RTI | 6,140 | 75 | 1.22 | (0.98, 1.53) | 3,140 | 75 | 2.39 | (1.91, 2.98) |
| Thromboembolic disease | 1,050 | 18 | 1.71 | (1.09, 2.69) | 1,113 | 21 | 1.89 | (1.24, 2.87) |

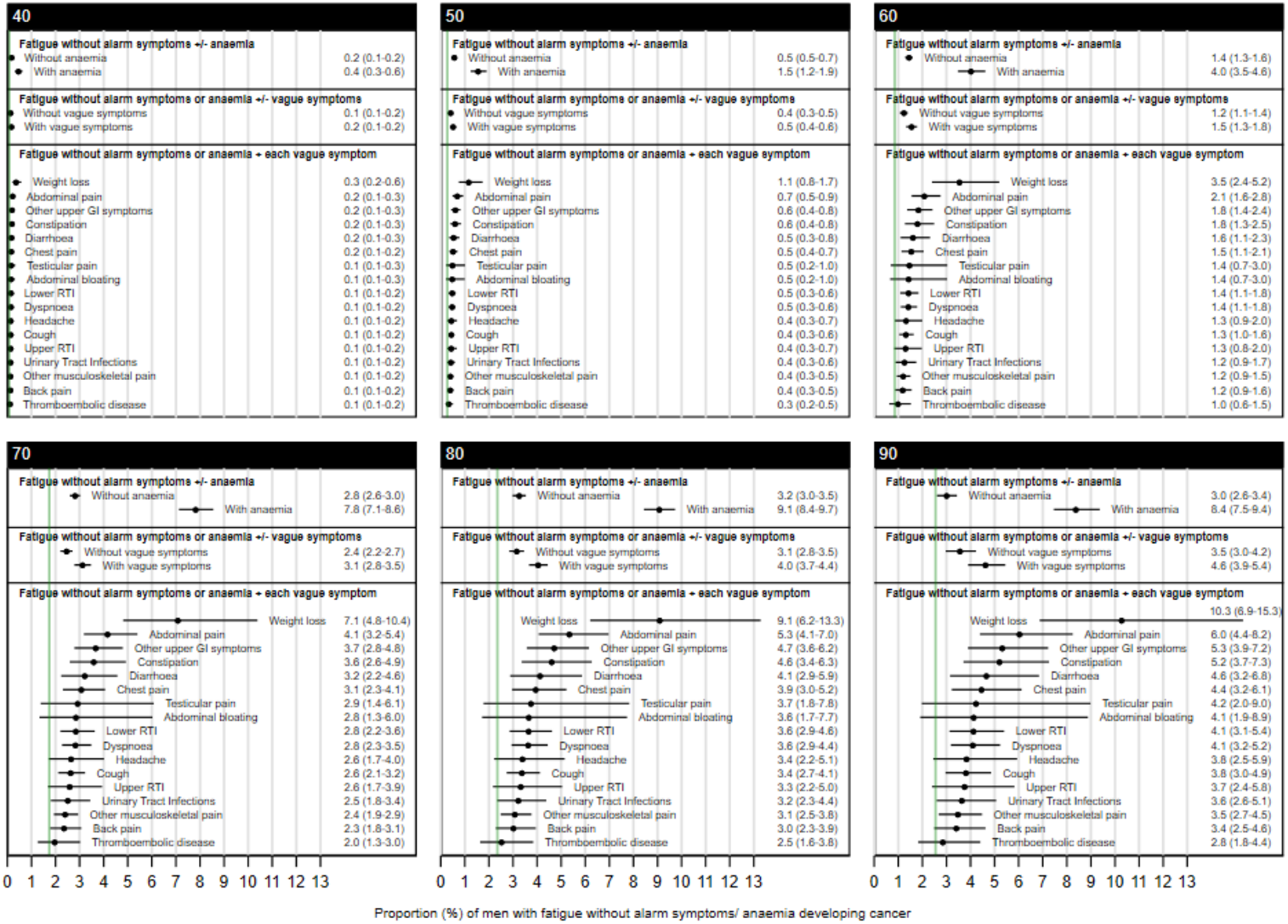



***Figure 5.5.Modelled nine-month cancer risk (%) in male patients with fatigue and no alarm symptoms, by presence of anaemia or each co-occurring vague symptom, for selected ages***

*Green line = observed nine-month cancer risk (%) for the general population in England in 2011, by five year age band. Available population estimates grouped all men aged 85+. Urinary Tract Infections also include cystitis, dysuria, urgency, painful urination, urine smell. Other upper GI (gastro-intestinal) symptom includes dyspepsia, nausea, vomiting, haematemesis, loss of appetite. RTI = Respiratory Tract Infections.*

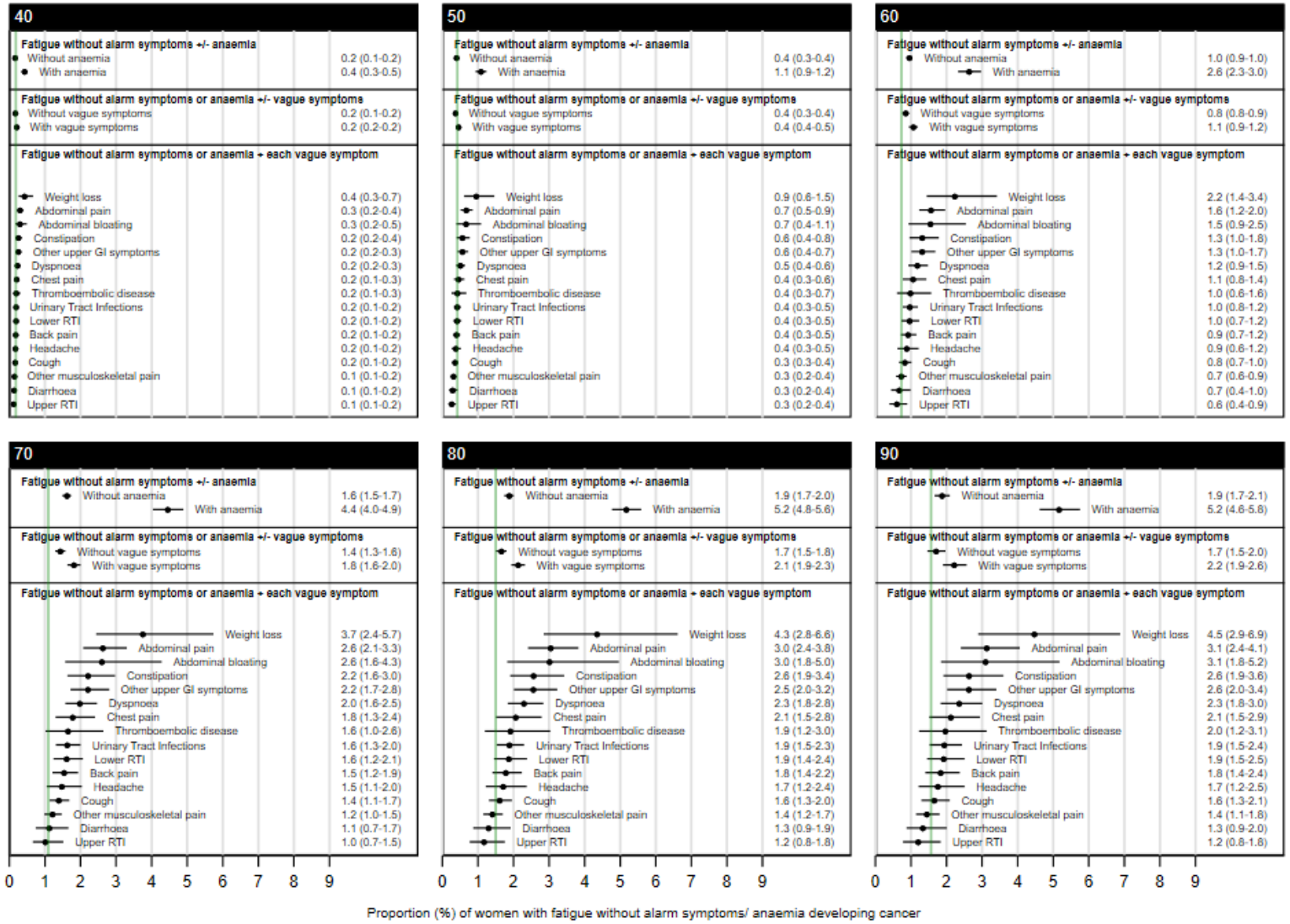



***Figure 5.6. Modelled nine-month cancer risk (%) in female patients with fatigue and no alarm symptoms, by presence of anaemia or each co-occurring vague symptom, for selected ages***

*Green line = observed nine-month cancer risk (%) for the general population in England in 2011, by five year age band. Available population estimates grouped all women aged 85+. Urinary Tract Infections also include cystitis, dysuria, urgency, painful urination, urine smell. Other upper GI (gastro-intestinal) symptom includes dyspepsia, nausea, vomiting, haematemesis, loss of appetite. RTI = Respiratory Tract Infections.*

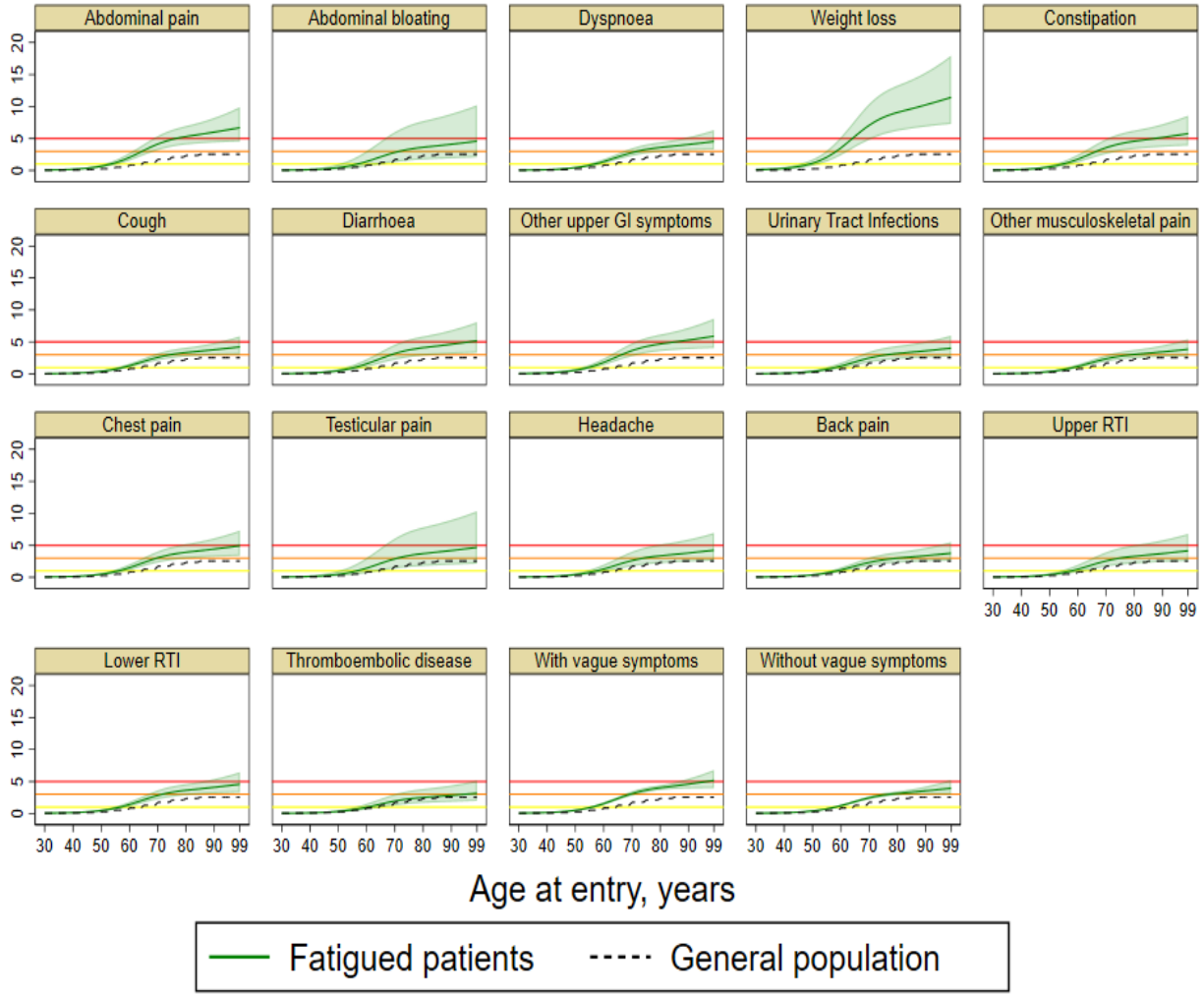



***Figure 5.7. Modelled cancer risk by year of age for co-occurring symptom combinations: men***

*Modelled nine-month cancer risk (%) in men with fatigue for each year of age (30-99), by symptom. Men with and without each vague symptom (restricted to patients with fatigue and no alarm symptom or anaemia). Risk for non-linear continuous age modelled using restricted cubic splines. Includes observed nine-month cancer risk (%) for the general population in England in 2011, by five year age band. Available population estimates grouped all men aged 85+. Urinary Tract Infections also include cystitis, dysuria, urgency, painful urination, urine smell. Other upper GI (gastro-intestinal) symptoms include dyspepsia, nausea, vomiting, haematemesis, loss of appetite. RTI = Respiratory Tract Infections.*

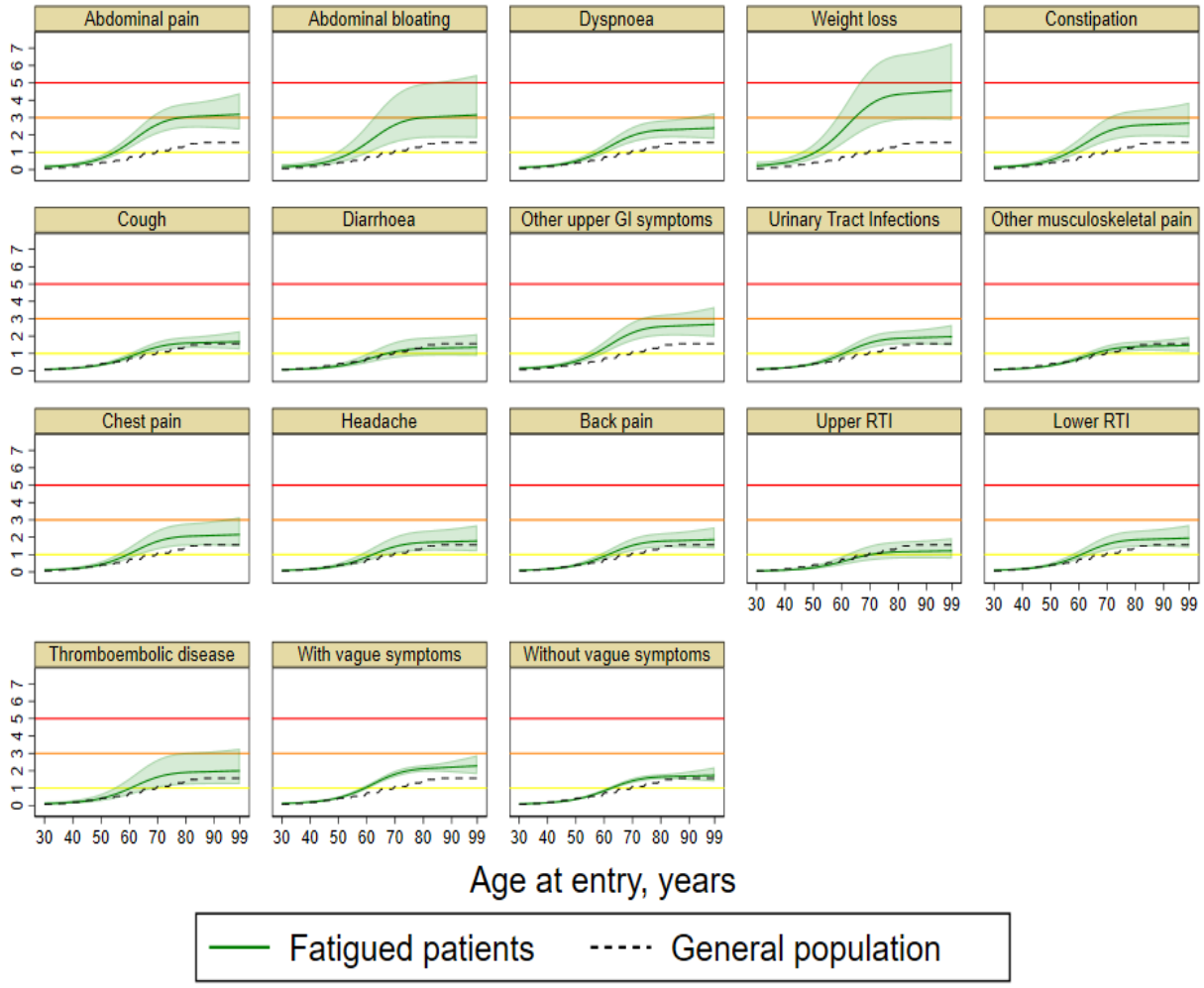



***Figure 5.8. Modelled cancer risk by year of age for co-occurring symptom combinations: women***

*Modelled nine-month cancer risk (%) in women with fatigue for each year of age (30-99), by symptom. Women with and without each vague symptom (restricted to patients with fatigue and no alarm symptom or anaemia). Risk for non-linear continuous age modelled using restricted cubic splines. Includes observed nine-month cancer risk (%) for the general population in England in 2011, by five year age band. Available population estimates grouped all women aged 85+. Urinary Tract Infections also include cystitis, dysuria, urgency, painful urination, urine smell. Other upper GI (gastro-intestinal) symptoms include dyspepsia, nausea, vomiting, haematemesis, loss of appetite. RTI = Respiratory Tract Infections.*

In men, the age at which risk exceeded 3% was 59 years (3.2% [2.2%-4.7%]) for fatigue-weight loss, 65 years (3.1% [2.4%-4.1%]) for fatigue-abdominal pain, 67 years (3.1% [2.2%-4.2%]) for fatigue-constipation, and 67 years (3.1% [2.4%-4.1%]) for fatigue-other upper GI symptoms. In women, risk exceeded 3% from 65 years (3.1% [2.0%-4.7%]) for fatigue-weight loss, and 79 years (3.0% [2.4%-3.8%]) for fatigue-abdominal pain, and 80 years for fatigue-abdominal bloating (3.0% [1.8%-5.0%]) (Table 5.3, Appendix 10.5.9).

***Table 5.3. Age (years) at which modelled nine-month cancer risk (%) exceeded 2%, 3%, and 6% in patients with fatigue without co-occurring alarm symptoms/ anaemia, by presence of each co-occurring vague symptom.***

| | Men | | | Women | | |
|---|---|---|---|---|---|---|
| | >2% | >3% | >6% | >2% | >3% | >6% |
| **Patients with fatigue, without alarm symptoms or anaemia** | | | | | | |
| With vague symptoms | 63 | 70 | . | 75 | - | - |
| Without vague symptoms | 67 | 78 | . | - | - | - |
| *Pairwise combinations of fatigue with each vague symptom:* | | | | | | |
| Abdominal bloating | 65 | 72 | . | 64 | 80 | . |
| Abdominal pain | 60 | 65 | 90 | 64 | 79 | . |
| Back pain | 67 | 80 | . | . | . | . |
| Chest pain | 64 | 70 | . | 77 | . | . |
| Constipation | 62 | 67 | . | 68 | . | . |
| Cough | 66 | 75 | . | . | . | . |
| Diarrhoea | 63 | 69 | . | . | . | . |
| Other upper GI symptoms | 61 | 67 | . | 68 | . | . |
| Dyspnoea | 65 | 72 | . | 71 | . | . |
| Headache | 66 | 74 | . | . | . | . |
| Lower RTI | 65 | 72 | . | . | . | . |
| Other musculoskeletal pain | 67 | 79 | . | . | . | . |
| Testicular pain | 64 | 71 | . | . | . | . |
| Thromboembolic disease | 71 | 95 | . | 99 | . | . |
| Upper RTI | 66 | 75 | . | . | . | . |
| Urinary Tract Infections | 66 | 77 | . | . | . | . |
| Weight loss | 55 | 59 | 67 | 59 | 65 | . |

*Urinary Tract Infections also include cystitis, dysuria, urgency, painful urination, urine smell. Other upper GI (gastro-intestinal) symptom includes dyspepsia, nausea, vomiting, haematemesis, loss of appetite. RTI = Respiratory Tract Infections.*

### 5.7.6 Sensitivity analyses

In the main analysis, co-occurring symptoms were identified if recorded three months before to one month after the patient's first fatigue presentation. In sensitivity analysis, broadening the look back time window to 12 months before fatigue presentation resulted in substantial increases in the proportions of fatigue presenters with both accompanying alarm symptoms, and accompanying vague symptoms (Table 5.4). This resulted in slightly lower risk of cancer, consistently across all symptom combinations examined (Table 5.5 and Table 5.6).

***Table 5.4. Frequency of co-occurring symptoms by time window used***

*Proportion (%) of patients with fatigue who had each co-occurring alarm or vague symptom, by time window used to define co-occurrence (time before/ after the first fatigue presentation (e.g. -3 months before the first fatigue presentation to +1 month after)). Cell counts under 5 are suppressed to reduce statistical disclosure risk. Urinary Tract Infections also include cystitis, dysuria, urgency, painful urination, urine smell. Other upper GI (gastro-intestinal) symptoms include dyspepsia, nausea, vomiting, haematemesis, loss of appetite. RTI = Respiratory Tract Infections.*

| | Same day | | -3 months/ +1 month | | -6 months/ +1 month | | -9 months/ +1 month | | -12 months/ +1 month | |
|---|---|---|---|---|---|---|---|---|---|---|
| | n | % | n | % | n | % | n | % | n | % |
| **All patients with fatigue** | **285,382** | | **285,382** | | **285,382** | | **285,382** | | **285,382** | |
| With alarm symptoms (excluding anaemia) | 2,957 | 1.0 | 10,380 | 3.6 | 15,193 | 5.3 | 19,998 | 7.0 | 24,614 | 8.6 |
| With anaemia | 5,577 | 2.0 | 37,082 | 13.0 | 41,560 | 14.6 | 45,286 | 15.9 | 48,654 | 17.0 |
| With vague symptoms | 33,231 | 11.6 | 113,207 | 39.7 | 147,700 | 51.8 | 172,011 | 60.3 | 189,925 | 66.6 |
| *Each alarm symptom* | | | | | | | | | | |
| Breast lump | 355 | 0.1 | 1,237 | 0.4 | 1,971 | 0.7 | 2,720 | 1.0 | 3,495 | 1.2 |
| Change in bowel habit | 295 | 0.1 | 902 | 0.3 | 1,316 | 0.5 | 1,758 | 0.6 | 2,185 | 0.8 |
| Nipple changes of concern | 88 | 0.0 | 234 | 0.1 | 355 | 0.1 | 508 | 0.2 | 666 | 0.2 |
| Dysphagia | 208 | 0.1 | 894 | 0.3 | 1,292 | 0.5 | 1,688 | 0.6 | 2,074 | 0.7 |
| Haematuria | 157 | 0.1 | 1,148 | 0.4 | 1,690 | 0.6 | 2,214 | 0.8 | 2,718 | 1.0 |
| Haemoptysis | 68 | 0.0 | 347 | 0.1 | 501 | 0.2 | 641 | 0.2 | 789 | 0.3 |
| Jaundice | 39 | 0.0 | 194 | 0.1 | 232 | 0.1 | 261 | 0.1 | 291 | 0.1 |
| Post-menopausal bleeding | 76 | 0.0 | 291 | 0.1 | 479 | 0.2 | 647 | 0.2 | 794 | 0.3 |
| Rectal bleeding | 554 | 0.2 | 2,035 | 0.7 | 3,031 | 1.1 | 3,985 | 1.4 | 4,921 | 1.7 |
| Abdominal mass/ intestinal obstruction | 49 | 0.0 | 218 | 0.1 | 316 | 0.1 | 412 | 0.1 | 505 | 0.2 |
| Breast skin changes | < 5 | - | 16 | 0.0 | 27 | 0.0 | 36 | 0.0 | 45 | 0.0 |
| Lymphadenopathy | 183 | 0.1 | 523 | 0.2 | 711 | 0.2 | 903 | 0.3 | 1,049 | 0.4 |
| Hoarseness | 215 | 0.1 | 686 | 0.2 | 995 | 0.3 | 1,320 | 0.5 | 1,666 | 0.6 |
| Head or neck lump | 145 | 0.1 | 442 | 0.2 | 663 | 0.2 | 878 | 0.3 | 1,081 | 0.4 |
| Testicular enlargement/ lump | 57 | 0.0 | 179 | 0.1 | 267 | 0.1 | 352 | 0.1 | 455 | 0.2 |
| Other lump | 529 | 0.2 | 1,413 | 0.5 | 2,113 | 0.7 | 2,894 | 1.0 | 3,661 | 1.3 |

| *Each vague symptom* | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Abdominal pain | 2,346 | 0.8 | 11,683 | 4.1 | 17,751 | 6.2 | 23,445 | 8.2 | 28,843 | 10.1 |
| Abdominal bloating | 516 | 0.2 | 1,574 | 0.6 | 2,353 | 0.8 | 3,089 | 1.1 | 3,831 | 1.3 |
| Dyspnoea | 3,294 | 1.2 | 12,194 | 4.3 | 16,676 | 5.8 | 20,613 | 7.2 | 23,968 | 8.4 |
| Night sweats | 187 | 0.1 | 430 | 0.2 | 568 | 0.2 | 706 | 0.2 | 859 | 0.3 |
| Weight loss | 618 | 0.2 | 1,799 | 0.6 | 2,279 | 0.8 | 2,737 | 1.0 | 3,206 | 1.1 |
| Constipation | 1,118 | 0.4 | 5,103 | 1.8 | 7,433 | 2.6 | 9,471 | 3.3 | 11,480 | 4.0 |
| Cough | 4,762 | 1.7 | 22,652 | 7.9 | 33,840 | 11.9 | 43,448 | 15.2 | 51,844 | 18.2 |
| Diarrhoea | 1,302 | 0.5 | 5,872 | 2.1 | 8,596 | 3.0 | 11,246 | 3.9 | 13,801 | 4.8 |
| Pelvic pain | 18 | 0.0 | 104 | 0.0 | 170 | 0.1 | 249 | 0.1 | 320 | 0.1 |
| Other upper GI symptoms | 3,127 | 1.1 | 9,889 | 3.5 | 13,992 | 4.9 | 17,863 | 6.3 | 21,411 | 7.5 |
| Urinary Tract Infections | 2,118 | 0.7 | 13,666 | 4.8 | 19,589 | 6.9 | 24,746 | 8.7 | 29,377 | 10.3 |
| Other musculoskeletal pain | 7,726 | 2.7 | 25,905 | 9.1 | 38,956 | 13.7 | 50,842 | 17.8 | 61,447 | 21.5 |
| Chest pain | 1,847 | 0.6 | 7,407 | 2.6 | 11,136 | 3.9 | 14,527 | 5.1 | 17,858 | 6.3 |
| Testicular pain | 167 | 0.1 | 565 | 0.2 | 854 | 0.3 | 1,165 | 0.4 | 1,509 | 0.5 |
| Headache | 2,920 | 1.0 | 9,377 | 3.3 | 13,628 | 4.8 | 17,696 | 6.2 | 21,586 | 7.6 |
| Back pain | 4,196 | 1.5 | 16,099 | 5.6 | 24,573 | 8.6 | 32,501 | 11.4 | 39,869 | 14.0 |
| Upper RTI | 1,025 | 0.4 | 7,441 | 2.6 | 12,118 | 4.2 | 16,400 | 5.7 | 20,374 | 7.1 |
| Lower RTI | 1,458 | 0.5 | 11,909 | 4.2 | 18,064 | 6.3 | 23,385 | 8.2 | 27,962 | 9.8 |
| Thromboembolic disease | 487 | 0.2 | 2,914 | 1.0 | 3,950 | 1.4 | 4,880 | 1.7 | 5,743 | 2.0 |

***Table 5.5. Sensitivity analysis of cancer risk by time window used to identify co-occurring symptoms: men***

*Sensitivity analysis showing impact on cancer risk of using a 12 month versus a 3 month lookback period to identify symptom co-occurrence. Observed nine-month cancer risk (%) for men with fatigue who had a co-occurring symptom 12 months/ 3 months before to 1 month after the first fatigue presentation, for a) all patients with fatigue, b) patients with fatigue without alarm symptoms, c) patients with fatigue without alarm symptoms or anaemia. Urinary Tract Infections also include cystitis, dysuria, urgency, painful urination, urine smell. Other upper GI (gastro-intestinal) symptoms include dyspepsia, nausea, vomiting, haematemesis, loss of appetite. RTI = Respiratory Tract Infections.*

| | 3 months before – 1 month after | | | | 12 months before – 1 month after | | | |
|---|---|---|---|---|---|---|---|---|
| | Total patients (N) Total | Patients with cancer (n) Total | % | (95% CI) | Total patients (N) Total | Patients with cancer (n) Total | % | (95% CI) |
| **a) All patients with fatigue** | **92,768** | **2,036** | **2.19** | **(2.10, 2.29)** | **92,768** | **2,036** | **2.19** | **(2.10, 2.29)** |
| With alarm symptoms | 3,464 | 194 | 5.60 | (4.88, 6.42) | 7,572 | 299 | 3.95 | (3.53, 4.41) |
| Breast lump | 51 | < 5 | - | - | 128 | < 5 | - | - |
| Change in bowel habit | 338 | 18 | 5.33 | (3.39, 8.26) | 796 | 35 | 4.40 | (3.18, 6.05) |
| Nipple changes of concern | 22 | < 5 | - | - | 49 | < 5 | - | - |
| Dysphagia | 367 | 22 | 5.99 | (3.99, 8.91) | 783 | 28 | 3.58 | (2.49, 5.12) |
| Haematuria | 572 | 35 | 6.12 | (4.43, 8.39) | 1,361 | 65 | 4.78 | (3.76, 6.04) |
| Haemoptysis | 180 | 15 | 8.33 | (5.11, 13.29) | 382 | 23 | 6.02 | (4.05, 8.87) |
| Jaundice | 111 | 27 | 24.32 | (17.29, 33.08) | 156 | 28 | 17.95 | (12.72, 24.72) |
| Post-menopausal bleeding | | | | | | | | |
| Rectal bleeding | 842 | 24 | 2.85 | (1.92, 4.21) | 1,902 | 55 | 2.89 | (2.23, 3.74) |
| Abdominal mass/ intestinal obstruction | 73 | 14 | 19.18 | (11.78, 29.66) | 164 | 15 | 9.15 | (5.62, 14.54) |
| Breast skin changes | | | | | | | | |
| Lymphadenopathy | 141 | 10 | 7.09 | (3.90, 12.56) | 262 | 12 | 4.58 | (2.64, 7.83) |
| Hoarseness | 202 | 11 | 5.45 | (3.07, 9.49) | 451 | 20 | 4.43 | (2.89, 6.75) |
| Head or neck lump | 111 | 12 | 10.81 | (6.29, 17.95) | 256 | 14 | 5.47 | (3.29, 8.97) |
| Testicular enlargement/ lump | 179 | 6 | 3.35 | (1.55, 7.12) | 455 | 11 | 2.42 | (1.36, 4.28) |
| Other lump | 397 | 11 | 2.77 | (1.55, 4.89) | 950 | 23 | 2.42 | (1.62, 3.61) |
| **b) Patients with fatigue without alarm symptoms** | **89,304** | **1,842** | **2.06** | **(1.97, 2.16)** | **85,196** | **1,737** | **2.04** | **(1.95, 2.14)** |
| With anaemia | 10,833 | 778 | 7.18 | (6.71, 7.68) | 12,618 | 808 | 6.40 | (5.99, 6.84) |
| **c) Patients with fatigue without alarm symptoms or anaemia** | **78,471** | **1,064** | **1.36** | **(1.28, 1.44)** | **72,578** | **929** | **1.28** | **(1.20, 1.36)** |
| With vague symptoms | 28,528 | 519 | 1.82 | (1.67, 1.98) | 43,771 | 643 | 1.47 | (1.36, 1.59) |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Without vague symptoms | 49,943 | 545 | 1.09 | (1.00, 1.19) | 28,807 | 286 | 0.99 | (0.88, 1.11) |
| *Pairwise combinations of fatigue with each vague symptom:* | | | | | | | | |
| Abdominal pain | 2,292 | 61 | 2.66 | (2.08, 3.40) | 4,988 | 90 | 1.80 | (1.47, 2.21) |
| Abdominal bloating | 308 | 7 | 2.27 | (1.11, 4.62) | 631 | 9 | 1.43 | (0.75, 2.69) |
| Dyspnoea | 3,632 | 99 | 2.73 | (2.24, 3.31) | 5,837 | 144 | 2.47 | (2.10, 2.90) |
| Night sweats | 113 | < 5 | - | - | 166 | < 5 | - | - |
| Weight loss | 439 | 27 | 6.15 | (4.26, 8.80) | 649 | 28 | 4.31 | (3.00, 6.16) |
| Constipation | 1,032 | 41 | 3.97 | (2.94, 5.35) | 1,905 | 47 | 2.47 | (1.86, 3.27) |
| Cough | 5,950 | 116 | 1.95 | (1.63, 2.33) | 11,705 | 206 | 1.76 | (1.54, 2.01) |
| Diarrhoea | 1,370 | 32 | 2.34 | (1.66, 3.28) | 2,794 | 40 | 1.43 | (1.05, 1.94) |
| Pelvic pain | 26 | < 5 | - | - | 77 | < 5 | - | - |
| Other upper GI symptoms | 2,493 | 62 | 2.49 | (1.94, 3.18) | 4,570 | 77 | 1.68 | (1.35, 2.10) |
| Urinary Tract Infections | 1,605 | 41 | 2.55 | (1.89, 3.45) | 2,704 | 55 | 2.03 | (1.57, 2.64) |
| Other musculoskeletal pain | 6,462 | 99 | 1.53 | (1.26, 1.86) | 13,542 | 180 | 1.33 | (1.15, 1.54) |
| Chest pain | 2,478 | 52 | 2.10 | (1.60, 2.74) | 4,964 | 82 | 1.65 | (1.33, 2.05) |
| Testicular pain | 463 | 7 | 1.51 | (0.73, 3.09) | 1,102 | 15 | 1.36 | (0.83, 2.23) |
| Headache | 1,928 | 23 | 1.19 | (0.80, 1.78) | 3,771 | 43 | 1.14 | (0.85, 1.53) |
| Back pain | 3,935 | 55 | 1.40 | (1.08, 1.81) | 8,766 | 106 | 1.21 | (1.00, 1.46) |
| Upper RTI | 1,583 | 23 | 1.45 | (0.97, 2.17) | 3,840 | 41 | 1.07 | (0.79, 1.45) |
| Lower RTI | 3,140 | 75 | 2.39 | (1.91, 2.98) | 6,340 | 113 | 1.78 | (1.48, 2.14) |
| Thromboembolic disease | 1,113 | 21 | 1.89 | (1.24, 2.87) | 1,806 | 34 | 1.88 | (1.35, 2.62) |

***Table 5.6. Sensitivity analysis of cancer risk by time window used to identify co-occurring symptoms: women***

*Sensitivity analysis showing impact on cancer risk of using a 12 month versus a 3 month lookback period to identify symptom co-occurrence. Observed nine-month cancer risk (%) for women with fatigue who had a co-occurring symptom 12 months/ 3 months before to 1 month after the first fatigue presentation, for a) all patients with fatigue, b) patients with fatigue without alarm symptoms, c) patients with fatigue without alarm symptoms or anaemia. Urinary Tract Infections also include cystitis, dysuria, urgency, painful urination, urine smell. Other upper GI (gastro-intestinal) symptoms include dyspepsia, nausea, vomiting, haematemesis, loss of appetite. RTI = Respiratory Tract Infections.*

| | 3 months before – 1 month after | | | | 12 months before – 1 month after | | | |
|---|---|---|---|---|---|---|---|---|
| | Total patients (N) Total | Patients with cancer (n) Total | % | (95% CI) | Total patients (N) Total | Patients with cancer (n) Total | % | (95% CI) |
| **a) All patients with fatigue** | **192,614** | **2,101** | **1.09** | **(1.05, 1.14)** | **192,614** | **2,101** | **1.09** | **(1.05, 1.14)** |
| With alarm symptoms | 6,916 | 208 | 3.01 | (2.63, 3.44) | 17,042 | 305 | 1.79 | (1.60, 2.00) |
| Breast lump | 1,186 | 50 | 4.22 | (3.21, 5.52) | 3,367 | 59 | 1.75 | (1.36, 2.25) |
| Change in bowel habit | 564 | 24 | 4.26 | (2.88, 6.25) | 1,389 | 41 | 2.95 | (2.18, 3.98) |
| Nipple changes of concern | 212 | < 5 | - | - | 617 | 5 | 0.81 | (0.35, 1.88) |
| Dysphagia | 527 | 18 | 3.42 | (2.17, 5.33) | 1,291 | 29 | 2.25 | (1.57, 3.21) |
| Haematuria | 576 | 14 | 2.43 | (1.45, 4.04) | 1,357 | 27 | 1.99 | (1.37, 2.88) |
| Haemoptysis | 167 | 7 | 4.19 | (2.05, 8.40) | 407 | 8 | 1.97 | (1.00, 3.83) |
| Jaundice | 83 | 22 | 26.51 | (18.20, 36.89) | 135 | 22 | 16.30 | (11.02, 23.44) |
| Post-menopausal bleeding | 291 | 10 | 3.44 | (1.88, 6.21) | 794 | 15 | 1.89 | (1.15, 3.09) |
| Rectal bleeding | 1,193 | 26 | 2.18 | (1.49, 3.17) | 3,019 | 48 | 1.59 | (1.20, 2.10) |
| Abdominal mass/ intestinal obstruction | 145 | 17 | 11.72 | (7.45, 17.97) | 341 | 19 | 5.57 | (3.60, 8.54) |
| Breast skin changes | 16 | < 5 | - | - | 45 | < 5 | - | - |
| Lymphadenopathy | 382 | 10 | 2.62 | (1.43, 4.75) | 787 | 12 | 1.52 | (0.87, 2.65) |
| Hoarseness | 484 | 7 | 1.45 | (0.70, 2.95) | 1,215 | 12 | 0.99 | (0.57, 1.72) |
| Head or neck lump | 331 | < 5 | - | - | 825 | 11 | 1.33 | (0.75, 2.37) |
| Testicular enlargement/ lump | | | | | | | | |
| Other lump | 1,016 | 11 | 1.08 | (0.61, 1.93) | 2,711 | 27 | 1.00 | (0.69, 1.45) |
| **b) Patients with fatigue without alarm symptoms** | **185,698** | **1,893** | **1.02** | **(0.97, 1.07)** | **175,572** | **1,796** | **1.02** | **(0.98, 1.07)** |
| With anaemia | 24,323 | 661 | 2.72 | (2.52, 2.93) | 30,477 | 696 | 2.28 | (2.12, 2.46) |
| **c) Patients with fatigue without alarm symptoms or anaemia** | **161,375** | **1,232** | **0.76** | **(0.72, 0.81)** | **145,095** | **1,100** | **0.76** | **(0.71, 0.80)** |
| With vague symptoms | 62,300 | 619 | 0.99 | (0.92, 1.07) | 95,152 | 800 | 0.84 | (0.78, 0.90) |
| Without vague symptoms | 99,075 | 613 | 0.62 | (0.57, 0.67) | 49,943 | 300 | 0.60 | (0.54, 0.67) |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Pairwise combinations of fatigue with each vague symptom:* | | | | | | | | |
| Abdominal pain | 6,644 | 96 | 1.44 | (1.18, 1.76) | 14,705 | 134 | 0.91 | (0.77, 1.08) |
| Abdominal bloating | 893 | 16 | 1.79 | (1.11, 2.89) | 1,969 | 20 | 1.02 | (0.66, 1.56) |
| Dyspnoea | 5,314 | 94 | 1.77 | (1.45, 2.16) | 9,303 | 131 | 1.41 | (1.19, 1.67) |
| Night sweats | 220 | < 5 | - | - | 421 | < 5 | - | - |
| Weight loss | 665 | 22 | 3.31 | (2.19, 4.96) | 1,098 | 30 | 2.73 | (1.92, 3.87) |
| Constipation | 2,432 | 48 | 1.97 | (1.49, 2.61) | 4,673 | 79 | 1.69 | (1.36, 2.10) |
| Cough | 12,237 | 118 | 0.96 | (0.81, 1.15) | 25,239 | 250 | 0.99 | (0.88, 1.12) |
| Diarrhoea | 2,816 | 26 | 0.92 | (0.63, 1.35) | 5,897 | 40 | 0.68 | (0.50, 0.92) |
| Pelvic pain | 56 | < 5 | - | - | 145 | < 5 | - | - |
| Other upper GI symptoms | 4,895 | 79 | 1.61 | (1.30, 2.01) | 9,494 | 105 | 1.11 | (0.91, 1.34) |
| Urinary Tract Infections | 8,664 | 104 | 1.20 | (0.99, 1.45) | 16,677 | 174 | 1.04 | (0.90, 1.21) |
| Other musculoskeletal pain | 14,700 | 116 | 0.79 | (0.66, 0.95) | 30,838 | 240 | 0.78 | (0.69, 0.88) |
| Chest pain | 3,420 | 45 | 1.32 | (0.98, 1.76) | 7,514 | 73 | 0.97 | (0.77, 1.22) |
| Testicular pain | | | | | | | | |
| Headache | 5,996 | 38 | 0.63 | (0.46, 0.87) | 12,383 | 66 | 0.53 | (0.42, 0.68) |
| Back pain | 9,153 | 84 | 0.92 | (0.74, 1.13) | 20,086 | 155 | 0.77 | (0.66, 0.90) |
| Upper RTI | 4,599 | 24 | 0.52 | (0.35, 0.78) | 11,182 | 57 | 0.51 | (0.39, 0.66) |
| Lower RTI | 6,140 | 75 | 1.22 | (0.98, 1.53) | 12,971 | 129 | 0.99 | (0.84, 1.18) |
| Thromboembolic disease | 1,050 | 18 | 1.71 | (1.09, 2.69) | 1,741 | 29 | 1.67 | (1.16, 2.38) |

## 5.8 Discussion

### 5.8.1 Summary

In patients presenting to primary care with fatigue without alarm symptoms or anaemia, I characterise the frequency of 19 co-occurring vague symptoms. Age-adjusted cancer risk was higher for those with any vague symptom studied, including four symptoms in men, and six in women. Cancer risk exceeded 3% in older men with fatigue and any vague symptom, reaching this threshold earliest for fatigue-weight loss (59 years), fatigue-abdominal pain (65 years), fatigue-constipation (67 years), and fatigue-other upper GI symptoms (67 years). For women, risk exceeded 3% only in older women with fatigue-weight loss (65 years), fatigue-abdominal pain (79 years), and fatigue-abdominal bloating (80 years).

### 5.8.2 Strengths and limitations

This study has a number of strengths. It uses high quality electronic health records from CPRD, which are broadly representative of the age, sex, and ethnicity distribution of the UK population(145). Linkage to 'gold standard' population-level cancer registration (NCRAS) data offered 'gold standard' ascertainment of cancer diagnoses(136).

Unlike most similar studies(36,38,44,45,49,110,164,174–184), the nine-month follow up for cancer was guided by previous evidence of the duration of increased cancer risk following first fatigue presentation(163). This study also demonstrated for the first time that cancer risk estimates would be lower if using longer look back periods before the first fatigue presentation for including co-occurring symptoms.

There are several limitations to this study. The study population is limited to patients who presented to primary care with fatigue and in whom their doctors deemed the symptom severe enough to be coded in their records(89), and does not represent the broader population of patients who experience fatigue in the community(26). Therefore, comparisons with the general population are intended only to contextualise risk(163).

GPs are more likely to code 'alarm' symptoms in patients' medical records when they suspect cancer(89). This could potentially inflate cancer risk estimates in cohorts of patients presenting with alarm symptoms, though it is unclear whether similar patterns (of selectively recording the symptom when cancer is suspected) also apply to fatigue. As I have also restricted the cohort in this study to patients with fatigue *without* alarm symptoms for cancer, this complicates interpretation of the potential effects of selective recording on the risks reported in my study. If some patients with fatigue also presented with an alarm symptom for cancer which was not recorded by the GP, then these patients would be included in my study, which could increase the reported cancer risk. However, if GPs did not record alarm symptoms as coded entries for these patients because they did not suspect cancer, then this could decrease the risk estimates.

I examined fatigue in combination with other potential cancer symptoms, where Read code lists were available for those symptoms. It is possible that a small minority of patients included in the

cohort of patients with fatigue and no alarm symptom had one of 12 alarm symptoms for which Read code lists were unavailable (Appendix 10.5.3), however, the symptoms that were not included are likely to occur rarely in practice. Future research could examine a wider range of alarm and vague symptoms using more recently available Read code lists(160) or lists developed in other coding systems(185).

Age and symptom-specific risk estimates were produced through the use of modelling. However, the number of patients with some co-occurring symptoms (e.g. abdominal bloating in men) and at some ages – especially age 90 and above – was small, resulting in imprecision of some age-symptom-specific risk estimates.

While not possible in this study due to sample size limitations, further stratification of exposures would be informative, for example, by morbidity status, the nature of co-occurring symptoms (e.g. chronic or recent onset), or by multiple combinations of symptoms (e.g. fatigue in combination with abdominal pain and abdominal bloating). Furthermore, I examined the risk of all cancers combined, whereas National Institute for Health and Care Excellence (NICE) Guidelines are usually based on the risk of a specific cancer site.

### 5.8.3 Comparison with literature

To my knowledge, this is the first study to characterise symptom co-occurrence in fatigue presenters, and to estimate cancer risk in patients with fatigue and a wide range of vague symptoms. Together with other evidence, the findings establish abdominal pain, weight loss and fatigue as vague symptoms that confer a substantial risk of cancer often exceeding normative risk thresholds, particularly in combination(44,45,48,49,165,166) Nevertheless, it should be noted that the mere presence of additional vague symptoms is a marker of elevated risk, particularly if two or more are present.

In addition, older men with fatigue-constipation or fatigue-other upper gastro-intestinal (GI) symptoms (which included dyspepsia, nausea, vomiting, haematemesis, loss of appetite) and older women with fatigue-abdominal bloating were also at elevated (>3%) risk of cancer. This is concordant with prior literature examining some of these abdominal symptoms either alone(44,45) or in combination with weight loss(49) or abdominal pain(48).

### 5.8.4 Implications for research and practice

My study illustrates the feasibility of producing cancer risk estimates for groups of patients with symptoms that co-occur with a vague symptom, such as fatigue. The detailed examination of cancer risk in patients presenting to primary care with new-onset fatigue in the absence of alarm symptoms for cancer can guide the management of a sizeable population of patients for whom diagnostic management is particularly challenging. My research shows that when patients present with fatigue without accompanying alarm symptoms or anaemia, cancer risk does not generally exceed current UK 3% referral threshold for urgent investigation for suspected cancer published by National Institute for Health and Care Excellence (NICE)(16) (except for men aged 73 and over). Nevertheless, in men, the presence of other vague symptoms, additional to fatigue, increases the risk of undiagnosed cancer to levels exceeding 3%. In older women, risk for certain combinations of vague

symptoms (fatigue-weight loss, fatigue-abdominal pain, and fatigue-abdominal bloating) also exceed these thresholds. These groups could be considered for inclusion in NICE referral guidelines for suspected cancer.

The risks of over investigation in older patients with non-specific symptoms must also be considered in referral recommendations. For instance, an 80-year old man presenting with fatigue and no alarm symptoms has over 3% risk of being diagnosed with cancer in the next year. Yet one-year cancer risk for men in the general population aged over 80 years already exceeds 3% (Chapter 4), so this could largely reflect his background risk (unrelated to new-onset fatigue). Applying a 3% referral threshold without consideration of a patient's baseline risk could lead to frequent, unwanted investigations, and negative mental and physical impacts for patients and increased healthcare service costs(19,186). Referral decisions need to weigh these risks against the benefits of ruling out serious physical disease such as cancer, incorporating patient preferences supported by accurate communication of diagnostic uncertainty(187). Further thought is also needed about whether urgent referral should be recommended if the risk of all cancers combined exceeds 3%, and its potential implications for over investigation, since the 3% threshold has so far generally applied to the risk of individual cancer sites in NICE referral guidelines (16).

My study examined patients with fatigue and other vague cancer symptoms. By their nature, vague symptoms are likely associated with a moderately raised risk of many different cancer sites. In my study, even the top three sites diagnosed following a fatigue-vague symptom combination typically excluded at least one third of cancers diagnosed. This varied mix of cancers also meant the ranking of cancer sites was not precise, as there were often several sites forming similar proportions of cancers diagnosed. In a different sample these could be expected to be ordered differently, hence it could not be said with certainty whether the risk of any one cancer site was higher than others. Future research should use larger sample sizes in order to adequately assess whether any vague symptoms, taken separately or in combination, could help to differentiate between the most likely cancer sites. It is possible that some of 'vague' non-alarm symptoms (as classified by I and clinical colleagues) are more organ-specific than others. Our list included urinary tract infections (UTIs), for example, as they are not included in NICE Guidelines as potential 'alarm' symptoms for cancer(16). In the case of abdominal pain and dyspepsia, previous research has shown that these symptoms, when considered separately, are associated with a diverse range of cancer sites(164). Combinations of vague symptoms could together further differentiate between the most likely cancer sites, particularly if incorporating other risk stratifiers (e.g. results of common blood tests). This is important in the UK context, as urgent 'two-week-wait' referral routes are cancer site-specific, being grouped within 16 medical specialties (e.g. haematology, lower GI etc.) (188), and non-specific diagnostic care pathways for patients with fatigue or other non-specific symptoms (Rapid Diagnostic Centres) (46), are not yet well established.

I also found that in patients with fatigue and no alarm symptoms, cancer risk exceeds 3% in patients with anaemia, rising to over 8% in older men. Although I have not characterised anaemia type (e.g. by iron-deficiency status), the findings indicate that low haemoglobin alongside fatigue confers a relatively high risk of cancer, which is supported by previous research into anaemia(108). Although fatigue can be directly attributable to anaemia, it is important that the risk of underlying cancer in these patients is also investigated, particularly in older patients. While existing NICE guidelines recommend appropriate investigation of anaemia, this alarm feature is not always appropriately investigated(189).

## 5.9 Additional information

### 5.9.1 Data availability

Data management and analysis code, and Read code lists, are available online at https://github.com/rmjlrwh/Fatigue.

### 5.9.2 Competing interests

MB receives personal fees from GRAIL Inc, for IDMC membership unrelated to this study. All other authors declare no competing interests.

## 5.10 Chapter summary

In this chapter, I established that fatigue is not usually recorded in primary care in combination with an alarm symptom for cancer. When studying only patients presenting with fatigue without accompanying alarm symptoms or anaemia, cancer risk did not exceed current UK National Institute for Health and Care Excellence (NICE) thresholds (> 3%) for urgent investigation for suspected cancer, except in men aged 73 years and over. However, risk did exceed referral thresholds in older men and women presenting with fatigue alongside other vague symptoms (in particular, weight loss, abdominal pain, abdominal bloating (women), constipation (men), or other upper gastrointestinal symptoms (men)). The age and sex specific risks reported can guide clinical decisions about referrals for specialist investigations for cancer, depending on the presence or absence of other vague symptoms presenting alongside fatigue.

# 6. Chapter 6: Risk of incident cancer compared to other diseases after presenting in primary care with fatigue: a population-based cohort study

## 6.1 Chapter rationale

In this chapter, I aimed to contextualise cancer risk in patients presenting with fatigue, by quantifying the risks of a wide range of other possible diagnoses, and how these risks vary by patient age and sex. This evidence could support UK diagnostic guidelines for suspected cancer and inform GPs about which diagnostic investigations to prioritise and referral pathways to consider when patients initially present with fatigue.

## 6.2 Publication

This chapter will be submitted to a peer reviewed journal (journal to be confirmed). For more information, see Appendix 10.6.1.

## 6.3 Author contributions

Authors: Becky White, Nadine Zakkak, Cristina Renzi, Meena Rafiq, Arturo Gonzalez-Izquierdo, Spiros Denaxas, Brian D Nicholson, Georgios Lyratzopoulos, Matthew Barclay

BW, MB, and GL conceived and designed the study. BW managed and analysed the data, under the supervision of MB, who also provided statistical expertise and designed the DAG diagrams in the discussion. MB and NZ shared analytical code used to manage and analyse the data, and NZ quality assured the final code and outputs. BDN, MR, CR, and GL provided clinical input, and AGI and SD developed disease phenotypes and advised on the presentation and discussion of results. All authors contributed to drafting and revising the article.

## 6.4 Abstract


**Background**

Fatigue is a non-localising symptom that is a feature of cancer but also many other diseases. Diagnostic guidelines recommending referral for suspected cancer in fatigue presenters currently lack context about the risk of non-neoplastic disease, yet this could support clinical decision-making.

**Methods**

I identified adults aged 30-99 years presenting with new-onset fatigue (fatigue presenters (FPs)) to English primary care 2007-2017, alongside a cohort of patients presenting without recorded fatigue (non-fatigue presenters (NFPs)). I described the excess short-term incidence of each of 237 diseases in FPs compared to NFPs. I modelled disease-specific 12-month risk by sex and calculated age-adjusted risk.

**Findings**

The study included 304,914 fatigue-presenters and 423,671 non-fatigue presenters. 127 of the 237 diseases studied were more common in male fatigue presenters than in male non-fatigue presenters, and 151 were more common in female fatigue presenters.

Age-adjusted cancer risk was higher among FPs than NFPs in men (FPs: 2.6% (CI = 2.5 to 2.7; NFPs: 1.2% (CI = 1.1 to 1.2; absolute excess risk (AER): 1.4%) and women (FPs: 1.4%, CI = 1.4 to 1.5; NFPs: 0.9%, CI = 0.9 to 0.9; AER: 0.5%). The relative frequency of cancer increased with age, particularly in men; by 80 years, cancer was the 3$^{rd}$ most common diagnosis and the disease with the 4$^{th}$ highest absolute excess risk in male fatigue presenters (male FPs: 7.0%, CI = 6.6 to 7.5; male NFPs: 3.4%, CI = 3.1 to 3.7); AER: 3.7%). In women, cancer remained relatively infrequent; by age 80 it was the disease with the 13$^{th}$ highest excess risk in fatigue presenters.

Overall, diseases that were most strongly associated with fatigue included: depression (e.g. male FPs: 3.2%, 95% C= 3.1 to 3.3; male NFPs: 0.8%, CI = 0.8 to 0.9; AER: 2.4%); insomnia & sleep disturbances (e.g. male FPs: 2.6%, CI = 2.5 to 2.7; male NFPs: 0.7%, CI = 0.6 to 0.8; AER: 1.9%); and hypo/hyperthyroidism in women only (female FPs: 2.4%, CI = 2.4 to 2.5; female NFPs: 0.7%, CI = 0.6 to 0.7; AER: 1.8%). Older men and women with fatigue were also at notably high actual risk (ranging from 2 to 5%) of consequential diagnoses including cancer, pneumonitis, acute kidney injury, stroke, chronic kidney disease, and coronary heart disease.

**Conclusions**

Among patients presenting to their GP with new-onset fatigue, I quantify and contextualise the risk of underlying cancer alongside the risk of a range of other diagnoses. By age 80 years in men, cancer is relatively more likely compared to other diagnoses, although several other consequential diseases should also be considered. In older women with fatigue, in the absence of other signs and symptoms of cancer, doctors could consider safety-netting for cancer or investigating it alongside other possible diagnoses. The findings highlight the importance of expanding multidisciplinary diagnostic services for patients with non-specific symptoms.

## 6.5 Background

Fatigue is a common presenting symptom in primary care and the principal complaint in an estimated 5–7% of consultations(21–24). Initial assessment of patients presenting with new-onset fatigue aims to rule out serious disease such as cancer by also considering the patient's age, sex and medical history, the presence of other presenting signs/ symptoms(37), and using first-line primary care diagnostic tests(54). However, the diagnosis is challenging, as fatigue is a non-specific symptom with low positive predictive value for a range of serious diseases (22,37,39,53).

In previous chapters, I quantified the short-term risk of cancer in patients with fatigue(163,190), but it is not known how this risk compares to other conditions. Fatigue can signal a range of other conditions, including but not limited to: coeliac disease, chronic fatigue syndrome, depression, hypothyroidism, iron deficiency, post-viral fatigue, and vitamin deficiency (22,37,39–42,53). More rarely, fatigue may indicate the presence of autoimmune disease (e.g. systemic lupus erythematosus), chronic infections (e.g. HIV, hepatitis C), heart disease or diabetes(37,53,54).

Existing UK diagnostic care guidance lists a range of potential diagnoses to consider in patients with fatigue (37,54). This is based largely on case-control studies of the prodromal features of specific diseases, since no population-level cohort study thus far has quantified the risk of multiple diagnostic outcomes in patients presenting with fatigue. Thus, the current guidance considers the risk of cancer in isolation, ignoring the risk of other possible diagnoses. A comprehensive evaluation of the risk of cancer relative to other diseases can help inform primary care clinicians about which diagnostic investigations to prioritise and referral pathways to consider.

To support general practitioners as they assess which serious and non-serious diagnoses to consider after initial presentation with new-onset fatigue, I aimed to quantify the short-term risk of cancer alongside a wide range of possible diagnoses, and how this varies by patient age and sex. To best inform diagnostic guidelines, I aimed to identify diseases with the strongest association with fatigue, by comparing risk in patients presenting with and without fatigue.

## 6.6 Methods

### 6.6.1 Study design and data source

I conducted a cohort study of patients with a record of fatigue presentation in primary care in England between 2007-2017, using electronic health records (EHRs) from the Clinical Practice Research Datalink (CPRD) GOLD (November 2021 database build). In 2013, the coverage of CPRD (GOLD) was 6.9% (N= 4.4 million) of the UK population(145). CPRD data items used include patients' recorded symptoms, demographic information (age, sex), and diagnoses in primary care.

Diagnoses recorded in secondary care from 1st April 1997 – 31st October 2020 were also identified through linkage with Hospital Episodes Statistics (HES) Admitted Patient Care (APC) data, a national dataset used for the National Health Service's (NHS) 'payment by results' reimbursement system and service planning. HES APC includes all inpatient admissions to NHS hospitals in England, which offers near to national coverage, since the majority of hospital activity in England is NHS-funded. It does not include attendances in accident and emergency or outpatients (i.e. those not requiring an inpatient bed)(191).

Cancers recorded in the national registry from 1st January 1995 - 31st December 2018 were also identified through linkage with the National Cancer Registration & Analysis Service (NCRAS), which offers complete ascertainment of cases in England(130). The linkages to HES APC and NCRAS data both used an eight-step deterministic linkage algorithm including NHS number, sex, date of birth, and postcode(192,193).

### 6.6.2 Fatigue presenter cohort

The fatigue presenter cohort was defined from a larger pool of patients previously identified in CPRD as presenting to primary care with a symptom of interest (including fatigue), between 1st January 2007- 31st December 2017, while aged 30-99 years. Fatigue was identified with a list of Read codes developed by WH and SP, using methods detailed by Watson et al(94) (listed in Appendix 10.6.2). The overall approach to identifying patients with new-onset fatigue is further detailed in a previous publication(163).

In short, I aimed to create a representative cohort of patients with 'new-onset' fatigue. Therefore, I excluded a small group of patients who had an 'ineligible' record of fatigue in the previous two years before their first eligible fatigue record (e.g. before the patient was 30 years old)). This ensured that patients did not enter the study midway through a series of (repeat) consultations for fatigue, and minimised the likelihood that fatigue was attributable to a previously diagnosed disease. To implement this exclusion criterion, and to later identify previous diagnoses, patients could only be considered for inclusion in the study if they had a fatigue record after they had been registered to an 'up-to-research-standard' CPRD practice for at least two years beforehand. Exclusion criteria are shown in Figure 6.1.

### 6.6.3 Comparison group cohorts

To illuminate the degree to which the observed disease risks among fatigue presenters were related specifically to fatigue, as opposed to simply feeling unwell enough to consult the GP, I examined disease incidence in 'non-fatigue presenters'. I identified a random sample of one million patients with at least one year of up to standard follow up in CPRD between 1st January 2007 – 31st October 2021 while aged 30-99 years, which was used to identify the non-fatigue presenters and registered patients. From this random sample, non-fatigue presenters were defined as patients with at least one consultation during a one year period (2011-2012), and without fatigue recorded in the previous two years or on the same date. A consultation was chosen at random to be the index date, provided there was no fatigue record in the previous two years.

In addition, it was theoretically possible that both fatigue and non-fatigue presenters were at greater or lower disease risk than the general population because of morbidity differences and increased disease severity in patients who present to primary care (a phenomenon sometimes termed the 'symptom iceberg') (113,114). Therefore, I further contextualised disease risk in fatigue presenters against the background risk in a second comparison group; the general population of 'registered patients'. This group was defined from the random sample of one million patients, as all patients with at least two years' follow up at a CPRD practice during the whole study period (2007-2017), while aged 30-99 years. As it aimed to estimate baseline risk regardless of healthcare seeking behaviour and consultation history, this group included all registered patients regardless of whether they consulted or had fatigue recorded (Figure 6.1).

### 6.6.4 Outcomes

I identified physical and mental health conditions using a large-scale disease phenotyping project by Kuan et al(170) (codes available at https://phenotypes.healthdatagateway.org), who included conditions that involve "intensive use of healthcare resources". These conditions are arranged within 14 broad groups defined by Kuan et al(170) (Appendix 10.6.3). The diseases of interest were supplemented with other fatigue-related conditions that were identified in NICE Guidelines and selected literature on fatigue(37,39,40,42,53,54). I then excluded conditions that were (a) irrelevant in the cohort of patients aged 30-99 years (e.g. congenital conditions which are diagnosed earlier in life), (b) could not represent incident disease (e.g. secondary malignancy), or (c) are usually a generic manifestation of another underlying disease (e.g. thrombocytosis)(Appendix 10.6.4). In total, I examined 237 diseases, including four diseases (chronic kidney disease, Lyme disease, pneumonitis, and insomnia & sleep disturbances) added from published codelists (104,170,188,194–198) other than Kuan et al.

The data sources (e.g. CPRD, HES APC, NCRAS) used by available phenotypes were predominantly used to define each disease in this study. An exception was made for malignant cancers; for these I only used 'gold standard' cancer registry data, as CPRD and HES APC may capture 'false positive' cases (133).

To ensure the disease phenotypes were not missing any common diagnosis codes, I searched for Read version 2 codes, ICD-10 codes, and Office of the Population Censuses and Surveys Classification of Interventions and Procedures version 4 (OPCS4) codes that were recorded in any of the data sources (CPRD Clinical, HES APC, NCRAS) in the three months after the first fatigue presentation in

fatigue-presenters. Codes occurring in more than 0.5% of patients were reviewed by clinicians, and 19 codes were added into the existing disease phenotypes. The full list of individual codes used in this study to define each phenotype are shown in Appendix 10.6.5.

I aimed to describe the risk of incident disease in patients who did not already have a previous diagnoses of that disease. As with similar previous studies(96,98,102,164), each outcome was analysed independently from the others; a 'disease-free' cohort was identified of patients with no previous diagnosis of that disease, and the risk of that disease estimated in that cohort. This meant that patients in one 'disease-free' cohort could have a previous diagnosis of another disease; for example, patients in the 'hypertension-free' cohort could have a previous diagnosis of diabetes. In addition, an individual patient could feature in more than one disease-free cohort and be counted as a case for more than one disease. In addition, some medical codes were categorised by Kuan et al under more than one disease(170). In the present study, the HES APC phenotypes contained the most duplicates; of the 1,973 ICD 10 codes used to define diseases, 1,674 were unique.

For most diseases, I excluded patients who ever had a record of that disease in primary or secondary care at any time before their index date, as a previous diagnosis could either signal the continued presence of a chronic disease (e.g. inflammatory bowel disease), or otherwise strongly influence the subsequent risk of that condition (e.g. cancer). For some infections that can have more than one incident occurrence during a life-course (e.g. UTI), I only excluded patients if they had a previous record of the infection within 2 years before their index date. In a sensitivity analysis, I examined the impact of including all patients, regardless of any previous diagnoses.

As discussed in a previous publication(156), it was appropriate to include diagnostic codes for chronic fatigue syndrome and postviral fatigue in the fatigue cohort, given I was identifying the patient's first new-onset fatigue record. However, when examining the risk of 'Postviral fatigue syndrome, neurasthenia and fibromyalgia' (which includes CFS), patients in the fatigue cohort who had a Read code for CFS or PVF recorded on their index date were excluded from the denominator, to avoid tautological identification of this outcome.

### 6.6.5 Follow up start

Follow up began with the patient's index date, which for fatigue presenters, was defined as the patient's first eligible record of fatigue during the study period (2007-2017). For non-fatigue presenters, it was a random presentation in the study period, provided there was no fatigue record in the previous two years. For registered patients, I selected a random day within the study period if it had at least two years follow up in CPRD before, regardless of whether the patient consulted on that day (Figure 6.1).

### 6.6.6 Follow up end

Follow up ended at the earliest of: twelve months following the index date, or the first diagnosis of the disease of interest. I chose twelve months as a compromise between capturing diagnoses of diseases that generally take a long time to diagnose (e.g. Parkinson's disease(189)), and the period of excess cancer risk following fatigue presentation, which has been shown to be up to nine

months(163). I included a supplementary analysis of monthly cumulative incidence of each disease, as the relative risk of different diseases could vary depending on the follow up time chosen(96).

I did not censor patients who died during the twelve month follow up period, and after death they remained in the denominator of patients 'at risk', because this approach would generate risk calculations that were more relevant to the way GPs assess risk at index presentation (115). I also did not censor patients if they left a CPRD practice during follow up, as changing GP practice is not random(116) and could introduce selection bias. However, diagnoses recorded after a patient died or left their practice were considered invalid, and these cases were excluded from the numerator.

### 6.6.7 Statistical analysis

I calculated the risk of each disease at 12 months following the index date. I modelled risk using Poisson regression models, stratified by sex, with disease diagnosis as the outcome and age modelled as a continuous exposure variable using natural cubic splines, and produced modelled disease risk at selected ages. Robust standard errors were used to account for possible overdispersion. Model fit for patients aged over 90 years was generally suboptimal, so estimates in this age range are not shown in results. I ascertained model fit for each cohort-sex-disease combination, by automatically flagging models where estimated risk in any single year of age (under 90 years) was at least three times higher or less than 20% of the observed risk in the respective 10-year age band. This flagged 296 of the total 1,422 models. Visual inspection showed that differences between estimated and observed values reflected the lack of granularity of using grouped age, rather than poor model fit. Disease-sex combinations that were impossible (e.g. male infertility in females) were excluded from results, and to ensure comparisons of estimates were sufficiently precise, diseases with fewer than 100 diagnosed fatigue presenters in each sex strata were also excluded.

For men and women at selected ages, I identified diseases with the greatest excess risk in patients with fatigue, by comparing modelled risk in fatigue presenters to that in non-fatigue presenters. The absolute excess risk was used to rank diseases for each sex and age combination.

I also compared the absolute risk of diseases overall in fatigue presenters against non-fatigue presenters and registered patients, stratified by sex. First, I calculated the overall unadjusted risk of each disease in fatigue presenters, for all ages combined. Then, because all-ages risk across diseases may be confounded by age, I age-standardised risk estimates in non-fatigue presenters and registered patients to the age distribution of fatigue presenters i.e. I calculated the expected risk in these two comparison groups, if their age profiles were the same as fatigue-presenters. To do this, I multiplied the total number of patients in each sex and year of age in fatigue presenters by the corresponding age-sex specific modelled disease risk in the respective comparison group, thereby obtaining the expected age- and sex-specific number of incident diseases. These were summed to calculate expected number of diagnoses for all men and all women combined in non-fatigue presenters and registered patients.

95% confidence intervals around the estimates for fatigue presenters and the comparison cohorts were used to determine whether differences were statistically significant.

Data management was conducted in MySQL Workbench version 6.1 and Stata version 17. All statistical analysis was conducted in R version 4.1.2 using the following key packages: stats::glm,

marginaleffects, splines, purrr, binom, tidyverse, dplyr, ggplot2. All analytical code is available at https://github.com/rmjlrwh/FatigueRiskMap. All code was quality assured by a second analyst (NZ) via independent replication of results for four randomly chosen diseases to ensure that code functioned as intended. This encompassed all code including data management, analysis, and generation of tables and figures. I used Strengthening the Reporting of Observational studies in Epidemiology (STROBE) guidelines for cohort studies (158) to report this study.

## 6.7 Results

### 6.7.1 Cohort inclusions and exclusions

402,975 patients were recorded in the Clinical Practice Research Datalink (CPRD) as presenting with fatigue between 2007-2017, of which 304,914 had at least one 'new-onset' fatigue presentation that met inclusion criteria (two years' follow up in CPRD before, while aged 30-99 years). For the non-fatigue presenters comparison group, of the random sample of 1 million patients in CPRD, 423,671 patients had at least one non-fatigue presentation between 2011-2012 that met inclusion criteria, without a fatigue presentation in the previous two years. For the registered patients comparison group, 759,904 patients of the 1 million random sample had at least two years follow up between 2007-2017, while aged 30-99 years (Figure 6.1).

After excluding patients with previous diagnoses of each disease before the index date in each of the three groups, the size of each sex-specific disease-free cohort ranged from 69,636 male fatigue presenters without previous hypertension (69% of male fatigue presenters), to 384,098 female registered patients without previous septicaemia (99.99% of female registered patients) (Appendix 10.6.6).

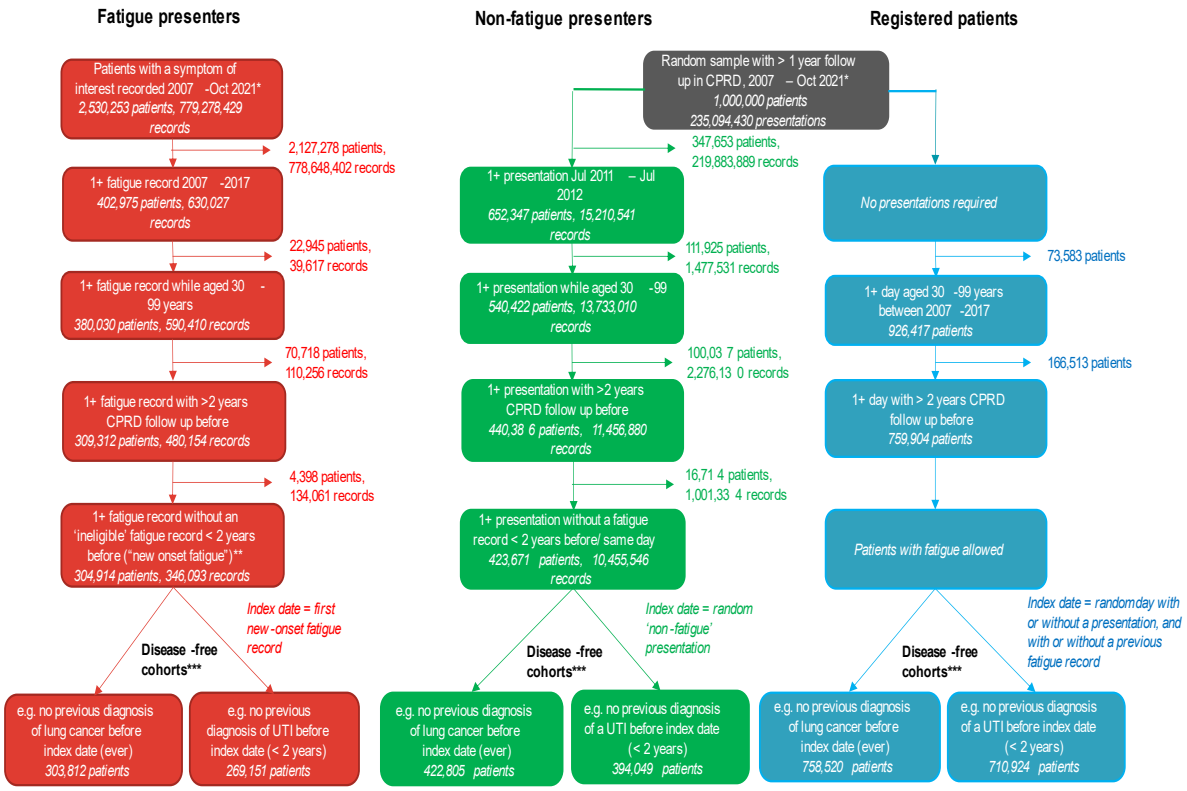



***Figure 6.1. Study cohorts***

**Data received from CPRD: Patients had to be aged 30-99 and within CPRD follow up when selected for the initial symptomatic/ random sample cohort. Follow up in CPRD began after the patient was registered to a CPRD practice and the practice's records were 'up to standard' (UTS) for research, i.e. the date from which the practice offered continuous service with no gaps in the recording of patient deaths or transfers. Follow up ended when the patient left the practice or died (if applicable), and when data was last collected from the practice.*

***'New-onset' fatigue: I excluded a small group of patients who had a prior 'ineligible' record of fatigue (e.g. before the patient was 30 years old) recorded shortly (<2 years) before their first 'eligible' record that met inclusion criteria. This ensured that when a patient entered the study with the their 'first' fatigue record, it was truly new-onset and did not begin midway through a series of consultations for fatigue.*

**** 'Disease-free' cohorts: An individual patient could feature in more than one 'disease-free' cohort and be counted as having the outcome for more than one disease. Illustrative disease-free cohorts are shown for the two-year lookback period for previous diagnoses of selected infections, and the lifetime lookback period for all other diseases. Sample sizes for all disease-free cohorts are included in supplementary appendices.*

### 6.7.2 Patient characteristics

Around two thirds (67%) of fatigue presenters (N=304,914) were women, which was higher than for non-fatigue presenters (54% women, N=423,671), and registered patients (51% women, N = 759,904). Among men, fatigue presenters tended to be older (median age 58 years) than non-fatigue presenters (55 years) or registered patients (49 years). A different pattern was observed in women: for fatigue presenters and registered patients the median age was 52 and 51 years respectively, but was slightly older in non-fatigue presenters (54 years).

For both men and women, fatigue presenters were slightly more likely to have a previous diagnosis of one of the diseases groups than non-fatigue presenters, and both groups were more likely than registered patients. For both men and women, the risk of a subsequent diagnosis of one of the diseases was higher in fatigue presenters compared to non-fatigue presenters, and both were higher than registered patients (Table 6.1).

### *Table 6.1.Demographic characteristics of study cohorts*

*Patients diagnosed previously or subsequently with each of the fourteen broad disease groups are given as a proportion of all fatigue presenters, non-fatigue presenters, or registered patients. The same patient could be diagnosed with more than one of the broad disease groups, so the number of patients diagnosed in each broad group do not equal the total number diagnosed with at least one of the studied diseases.*

| | Men | | | | | | Women | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Fatigue presenters | | Non-fatigue presenters | | Registered patients | | Fatigue presenters | | Non-fatigue presenters | | Registered patients | |
| | n | % | n | % | n | % | n | % | n | % | n | % |
| **Age group** | | | | | | | | | | | | |
| 30-39 | 13,313 | 13.2 | 32,142 | 16.38 | 102,064 | 27.16 | 42,246 | 20.71 | 40,415 | 17.77 | 99,980 | 26.03 |
| 40-49 | 19,285 | 19.12 | 44,183 | 22.52 | 87,195 | 23.2 | 47,523 | 23.29 | 50,312 | 22.12 | 80,084 | 20.85 |
| 50-59 | 20,142 | 19.97 | 40,463 | 20.63 | 69,985 | 18.62 | 36,931 | 18.1 | 43,472 | 19.11 | 66,349 | 17.27 |
| 60-69 | 19,534 | 19.36 | 40,347 | 20.57 | 57,440 | 15.28 | 28,867 | 14.15 | 41,641 | 18.3 | 57,259 | 14.91 |
| 70-79 | 16,838 | 16.69 | 25,245 | 12.87 | 36,743 | 9.78 | 26,477 | 12.98 | 28,725 | 12.63 | 41,586 | 10.83 |
| 80-89 | 10,336 | 10.25 | 11,971 | 6.1 | 19,245 | 5.12 | 18,235 | 8.94 | 18,307 | 8.05 | 30,248 | 7.87 |
| 90+ | 1,433 | 1.42 | 1,823 | 0.93 | 3,127 | 0.83 | 3,754 | 1.84 | 4,625 | 2.03 | 8,599 | 2.24 |
| **Median age** | | | | | | | | | | | | |
| Years | 58 | - | 55 | - | 49 | - | 52 | - | 54 | - | 51 | - |
| **Previous diagnoses** | | | | | | | | | | | | |
| Any disease studied | 92,502 | 91.69 | 164,835 | 84.02 | 274,199 | 72.96 | 193,244 | 94.71 | 203,744 | 89.56 | 325,963 | 84.86 |
| Benign Neoplasm/CIN | 5,260 | 5.21 | 7,169 | 3.65 | 10,220 | 2.72 | 22,544 | 11.05 | 21,637 | 9.51 | 31,371 | 8.17 |
| Cancers | 8,752 | 8.68 | 9,772 | 4.98 | 14,788 | 3.94 | 15,665 | 7.68 | 16,307 | 7.17 | 25,013 | 6.51 |
| Diseases of the Cardiovascular system | 38,437 | 38.1 | 55,499 | 28.29 | 78,301 | 20.84 | 52,924 | 25.94 | 55,784 | 24.52 | 80,927 | 21.07 |
| Diseases of the Circulatory System | 16,880 | 16.73 | 19,354 | 9.87 | 28,915 | 7.69 | 20,313 | 9.96 | 17,593 | 7.73 | 27,097 | 7.05 |
| Diseases of the Digestive System | 39,157 | 38.82 | 56,313 | 28.71 | 85,005 | 22.62 | 73,033 | 35.79 | 63,521 | 27.92 | 93,280 | 24.29 |
| Diseases of the Ear | 14,988 | 14.86 | 20,906 | 10.66 | 29,239 | 7.78 | 23,142 | 11.34 | 21,553 | 9.47 | 30,053 | 7.82 |
| Diseases of the Endocrine System | 14,436 | 14.31 | 19,354 | 9.87 | 26,559 | 7.07 | 37,353 | 18.31 | 32,349 | 14.22 | 48,247 | 12.56 |
| Diseases of the Eye | 16,528 | 16.38 | 23,093 | 11.77 | 32,606 | 8.68 | 29,728 | 14.57 | 29,748 | 13.08 | 43,781 | 11.4 |
| Diseases of the Genitourinary system | 36,543 | 36.22 | 50,257 | 25.62 | 72,800 | 19.37 | 79,463 | 38.95 | 74,018 | 32.54 | 108,916 | 28.36 |
| Diseases of the Respiratory System | 32,951 | 32.66 | 48,939 | 24.95 | 77,683 | 20.67 | 67,999 | 33.33 | 61,957 | 27.23 | 94,202 | 24.53 |
| Haematological/Immunological conditions | 10,782 | 10.69 | 10,855 | 5.53 | 16,058 | 4.27 | 35,461 | 17.38 | 27,560 | 12.11 | 41,397 | 10.78 |
| Infectious Diseases | 27,822 | 27.58 | 33,386 | 17.02 | 56,034 | 14.91 | 76,028 | 37.26 | 61,226 | 26.91 | 102,393 | 26.66 |
| Mental Health Disorders | 33,057 | 32.77 | 43,987 | 22.42 | 72,225 | 19.22 | 86,770 | 42.53 | 71,996 | 31.65 | 110,277 | 28.71 |
| Musculoskeletal conditions | 45,417 | 45.02 | 70,993 | 36.19 | 101,447 | 27 | 88,438 | 43.34 | 87,693 | 38.55 | 123,401 | 32.13 |
| Neurological conditions | 13,369 | 13.25 | 17,668 | 9.01 | 27,578 | 7.34 | 38,947 | 19.09 | 34,300 | 15.08 | 52,930 | 13.78 |
| Skin conditions | 33,884 | 33.59 | 54,052 | 27.55 | 83,046 | 22.1 | 76,676 | 37.58 | 75,110 | 33.02 | 110,225 | 28.7 |
| **Subsequent diagnoses** | | | | | | | | | | | | |
| Any disease studied | 64,018 | 63.46 | 78,945 | 40.24 | 117,421 | 31.25 | 131,853 | 64.62 | 104,520 | 45.94 | 157,894 | 41.11 |
| Benign Neoplasm/CIN | 1,390 | 1.38 | 1,529 | 0.78 | 2,150 | 0.57 | 3,958 | 1.94 | 2,854 | 1.25 | 4,308 | 1.12 |
| Cancers | 2,680 | 2.66 | 1,974 | 1.01 | 3,735 | 0.99 | 2,998 | 1.47 | 2,159 | 0.95 | 3,952 | 1.03 |
| Diseases of the Cardiovascular system | 18,651 | 18.49 | 17,787 | 9.07 | 27,465 | 7.31 | 22,657 | 11.1 | 17,478 | 7.68 | 27,637 | 7.2 |
| Diseases of the Circulatory System | 6,503 | 6.45 | 5,219 | 2.66 | 8,389 | 2.23 | 7,223 | 3.54 | 4,794 | 2.11 | 7,954 | 2.07 |
| Diseases of the Digestive System | 9,479 | 9.4 | 9,928 | 5.06 | 14,532 | 3.87 | 16,164 | 7.92 | 10,918 | 4.8 | 15,894 | 4.14 |
| Diseases of the Ear | 2,537 | 2.51 | 2,851 | 1.45 | 4,113 | 1.09 | 4,037 | 1.98 | 2,917 | 1.28 | 4,305 | 1.12 |
| Diseases of the Endocrine System | 5,088 | 5.04 | 3,975 | 2.03 | 5,734 | 1.53 | 14,731 | 7.22 | 8,303 | 3.65 | 12,771 | 3.32 |
| Diseases of the Eye | 4,722 | 4.68 | 5,967 | 3.04 | 8,539 | 2.27 | 7,507 | 3.68 | 6,529 | 2.87 | 9,850 | 2.56 |
| Diseases of the Genitourinary system | 12,745 | 12.63 | 11,500 | 5.86 | 18,636 | 4.96 | 20,761 | 10.18 | 13,722 | 6.03 | 22,097 | 5.75 |
| Diseases of the Respiratory System | 12,527 | 12.42 | 13,139 | 6.7 | 19,498 | 5.19 | 20,867 | 10.23 | 15,905 | 6.99 | 23,597 | 6.14 |
| Haematological/Immunological conditions | 5,295 | 5.25 | 2,863 | 1.46 | 4,307 | 1.15 | 12,695 | 6.22 | 4,906 | 2.16 | 8,034 | 2.09 |
| Infectious Diseases | 18,833 | 18.67 | 20,292 | 10.34 | 30,244 | 8.05 | 49,212 | 24.12 | 37,330 | 16.41 | 55,341 | 14.41 |
| Mental Health Disorders | 11,045 | 10.95 | 9,036 | 4.61 | 14,917 | 3.97 | 25,082 | 12.29 | 13,478 | 5.92 | 21,475 | 5.59 |
| Musculoskeletal conditions | 11,283 | 11.18 | 13,870 | 7.07 | 19,402 | 5.16 | 22,611 | 11.08 | 17,476 | 7.68 | 25,030 | 6.52 |
| Neurological conditions | 5,563 | 5.51 | 5,592 | 2.85 | 8,592 | 2.29 | 12,354 | 6.05 | 9,101 | 4 | 14,837 | 3.86 |
| Skin conditions | 7,276 | 7.21 | 9,176 | 4.68 | 13,001 | 3.46 | 14,868 | 7.29 | 11,359 | 4.99 | 16,734 | 4.36 |
| **Total** | | | | | | | | | | | | |
| N | 100,881 | - | 196,174 | - | 375,799 | - | 204,033 | - | 227,497 | - | 384,105 | - |

### 6.7.3 Age-adjusted risk

As illustrated by Figure 6.2 and Figure 6.3, disease risk was generally elevated in fatigue presenters compared to non-fatigue presenters. Among patients who were previously free of each of the 237 diseases studied, subsequent risk was higher in fatigue presenters compared to non-fatigue presenters for 127 diseases for men (of which, 121 were statistically significant), and 151 diseases for women (of which, 130 were statistically significant), after adjusting for differences in age. The other diseases (110 in men, 84 in women) were excluded from further analysis as there were under 100 patients with the outcome in fatigue presenters. Risk of 117 diseases was higher in both men and women (of which, 100 were statistically significant in both men and women). Disease risk was also higher in non-fatigue presenters than registered patients for 62 diseases in men (of which, 11 were statistically significant), and 66 diseases in women (of which, 12 were statistically significant), but was lower in 46 diseases in men (of which, 8 were statistically significant), and 57 in women (of which, 10 were statistically significant) (Appendix 10.6.7).

The risk of all cancers combined in men was 2.6% (CI = 2.5 to 2.7) in fatigue-presenters, which was approximately double that in non-fatigue presenters (1.2% (CI = 1.1 to 1.2)), making cancer the disease with the 7$^{th}$ greatest excess risk in fatigue presenters. In women, the risk of all cancers combined in fatigue presenters (1.4%, CI = 1.4 to 1.5), compared to (0.9%, CI = 0.9 to 0.9) in non-fatigue presenters, making cancer the disease with the 21$^{st}$ greatest (or 0.5 percentage points) excess risk in fatigue presenters. The five diseases with the greatest excess risk in fatigue-presenters were, in men: depression, lower RTIs, hypertension, insomnia & sleep disturbances, and ear and upper RTIs; and in women: depression, UTIs, ear and upper RTIs, hypo or hyperthyroidism, and lower RTIs. Absolute excess risk estimates for all 237 diseases are available in Appendix 10.6.7.

Regarding its absolute risk, in men, cancer ranked as the 7$^{th}$ most common disease in fatigue presenters, and the 8$^{th}$ most common in non-fatigue presenters. In women, cancer was the 19$^{th}$ most common disease in fatigue presenters, and the 13$^{th}$ most common in non-fatigue presenters (Appendix 10.6.6). The five most common new diagnoses in fatigue presenters were, in men: lower RTIs, hypertension, ear and upper RTIs, depression, UTIs (Figure 6.2); and in women: depression, UTIs, ear and upper RTIs, hypo or hyperthyroidism, and lower RTIs (Figure 6.3).

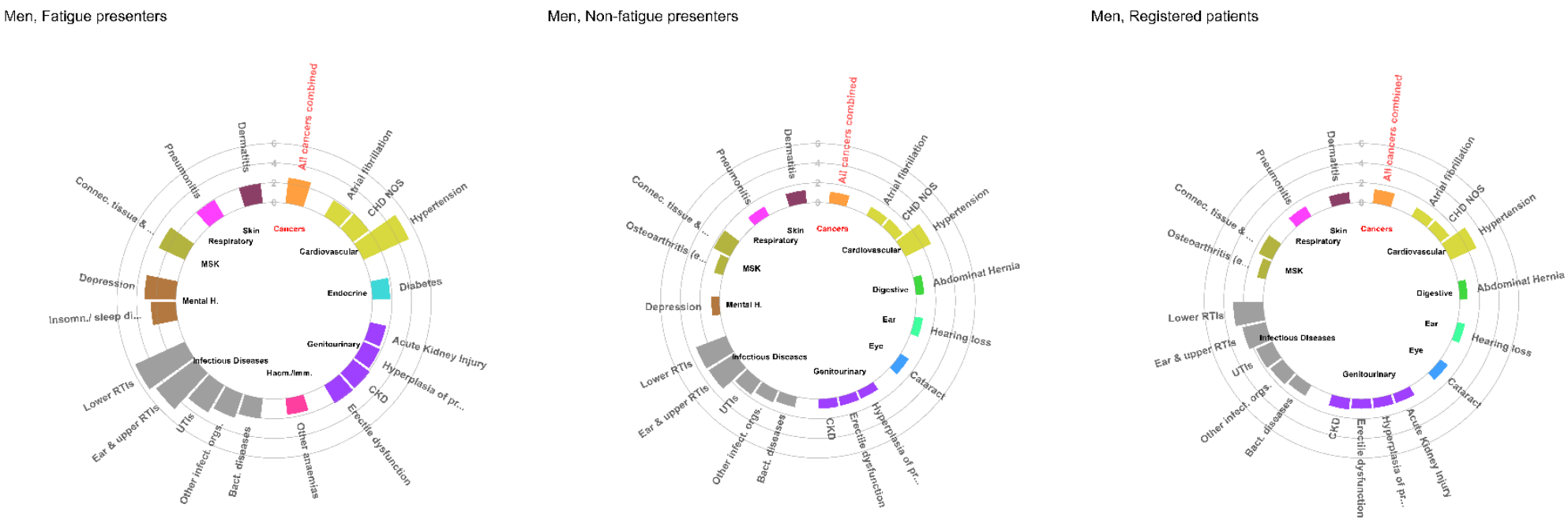



***Figure 6.2. Risk of 20 most common incident diseases in men presenting with new-onset fatigue, compared to non-fatigue presenters and in registered patients, after adjusting for age***

*The 20 most common diseases in each cohort are shown, including all cancers combined. See appendices for risk for all 237 diseases. Concentric circles demarcate risk (%) in increments of two percentage points. Each bar represents risk as a proportion of a different denominator; the cohort for each disease is different as patients with previous diagnoses of each disease are excluded. See appendices for relevant sensitivity analyses. For non-fatigue presenters and registered patients, I show the expected risk if the age profiles were the same as for fatigue-presenters.*

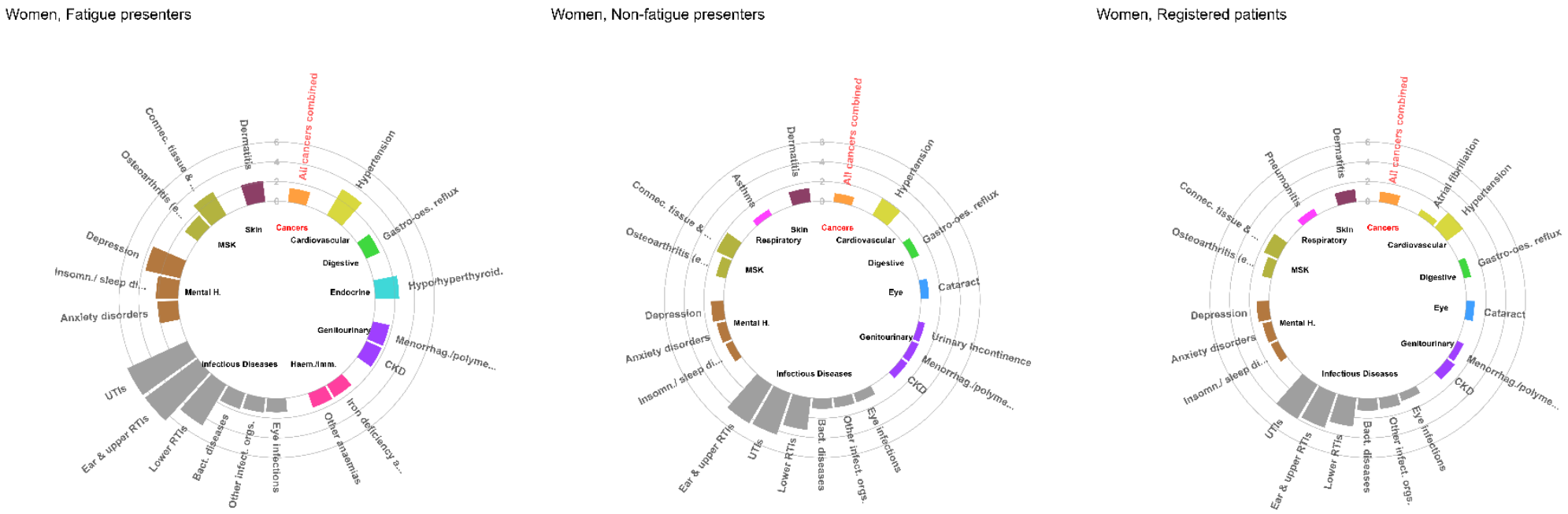



*Figure 6.3. Risk of 20 most common incident diseases in women presenting with new-onset fatigue, compared to non-fatigue presenters and in registered patients, after adjusting for age*

*The 20 most common diseases in each cohort are shown, including all cancers combined. See appendices for risk for all 237 diseases. Concentric circles demarcate risk (%) in increments of two percentage points. Each bar represents risk as a proportion of a different denominator; the cohort for each disease is different as patients with previous diagnoses of each disease are excluded. See appendices for relevant sensitivity analyses. For non-fatigue presenters and registered patients, I show the expected risk if the age profiles were the same as for fatigue-presenters.*

### 6.7.4 Age-specific excess risk

Excess risk of cancer in fatigue presenters increased with age. For men at age 40 years, cancer was the disease with the 35th greatest absolute excess risk in fatigue presenters (FPs) compared to non-fatigue presenters (NFPs) (AER = 1.3%; FPs: 0.23%, CI = 0.18 to 0.29 vs NFPs: 0.09%, CI = 0.07 to 0.11). By 60 years cancer had the 9th greatest excess risk (AER = 1.3%; FPs: 2.3%, CI = 2.1 to 2.5 vs NFPs: 1.0%, CI = 0.9 to 1.1), and by age 80 it had the 3rd (AER = 3.7%; 7.0%, CI = 6.6 to 7.5 vs 3.4%, CI = 3.1 to 3.7). By age 80, cancer was the 4th most common new diagnosis in male fatigue presenters (Figure 6.4, Appendix 10.6.8).

Excess risk of cancer in female fatigue presenters also increased with age, although age-specific excess risk was generally smaller than for men. For women at age 40 years, cancer was the disease with the 61st greatest absolute excess risk in FPs compared to NFPs (AER = 0.7%; 0.41%, CI = 0.36 to 0.46 vs 0.36%, CI = 0.32 to 0.41). By 60 years cancer had the 16th greatest excess risk (AER = 0.7%; 1.6%, CI = 1.5 to 1.7 vs 0.9%, CI = 0.8 to 1.0), and by 80 years it had the 13th (AER = 1.5%; 3.6%, CI = 3.4 to 3.9 vs 2.2%, CI = 2.0 to 2.4). By 80 years, cancer was also the 13th most common new diagnosis in female fatigue presenters (Figure 6.5, Appendix 10.6.8).

The diseases with the greatest excess risk in fatigue presenters compared to non-fatigue presenters varied by age for both men and women, with older patients (aged 80) being at higher excess risk of a different disease spectrum than those aged 40 or 60 years.

For men at age 40 years, the three diseases with the largest absolute excess risk (AER) in fatigue-presenters compared to non-fatigue presenters were depression (AER = 2.9%; 4.2%, CI = 3.9 to 4.5 vs 1.2%, CI = 1.1 to 1.4), ear and upper RTIs (AER = 1.9%; 5.2%, CI = 4.9 to 5.6 vs 3.4%, CI = 3.2 to 3.5), and insomnia & sleep disturbances (AER = 1.9%; 2.6%, CI = 2.4 to 2.9 vs 0.7%, CI = 0.7 to 0.8). By 80 years, the largest excess risk was for hypertension (AER = 4.5%; 12.7%, CI = 11.9 to 13.6 vs 8.2%, CI = 7.6 to 8.8), lower RTIs (AER = 4.0%; 10.1%, CI = 9.5 to 10.6 vs 6.0%, CI = 5.7 to 6.4), and all cancers combined (AER = 3.7%; 7.0%, CI = 6.6 to 7.5 vs 3.4%, CI = 3.1 to 3.7) (Figure 6.4).

For women at age 40 years, the three diseases with the largest absolute excess risk in fatigue-presenters compared to non-fatigue presenters were depression (AER = 2.7%; 4.5%, CI = 4.3 to 4.7 vs 1.8%, CI = 1.7 to 1.9), menorrhagia & polymenorrhoea (AER = 2.6%; 4.5%, CI = 4.3 to 4.8 vs 2.0, CI = 1.8 to 2.1), and ear and upper RTIs (AER = 2.4%; 7.7%, CI = 7.4 to 7.9 vs 5.2%, CI = 5.0 to 5.4). By 60 years, there was also large excess risk of hypo/hyperthyroidism (AER = 2.2%; 2.9%, CI = 2.7 to 3.1 vs 0.7%, CI =0.6 to 0.8). By 80 years, the largest excesses were in UTIs (AER = 4.0%; 11.4%, CI = 10.9 to 11.8 vs 7.3%, CI = 7.0 to 7.7), hypertension (AER = 3.9%; 11.7%, CI = 11.1 to 12.3 vs 7.8%, CI = 7.3 to 8.3), and chronic kidney disease (AER = 3.5%; 6.6%, CI = 6.2 to 6.9 vs 3.1%, CI = 2.9 to 3.4) (Figure 6.5).

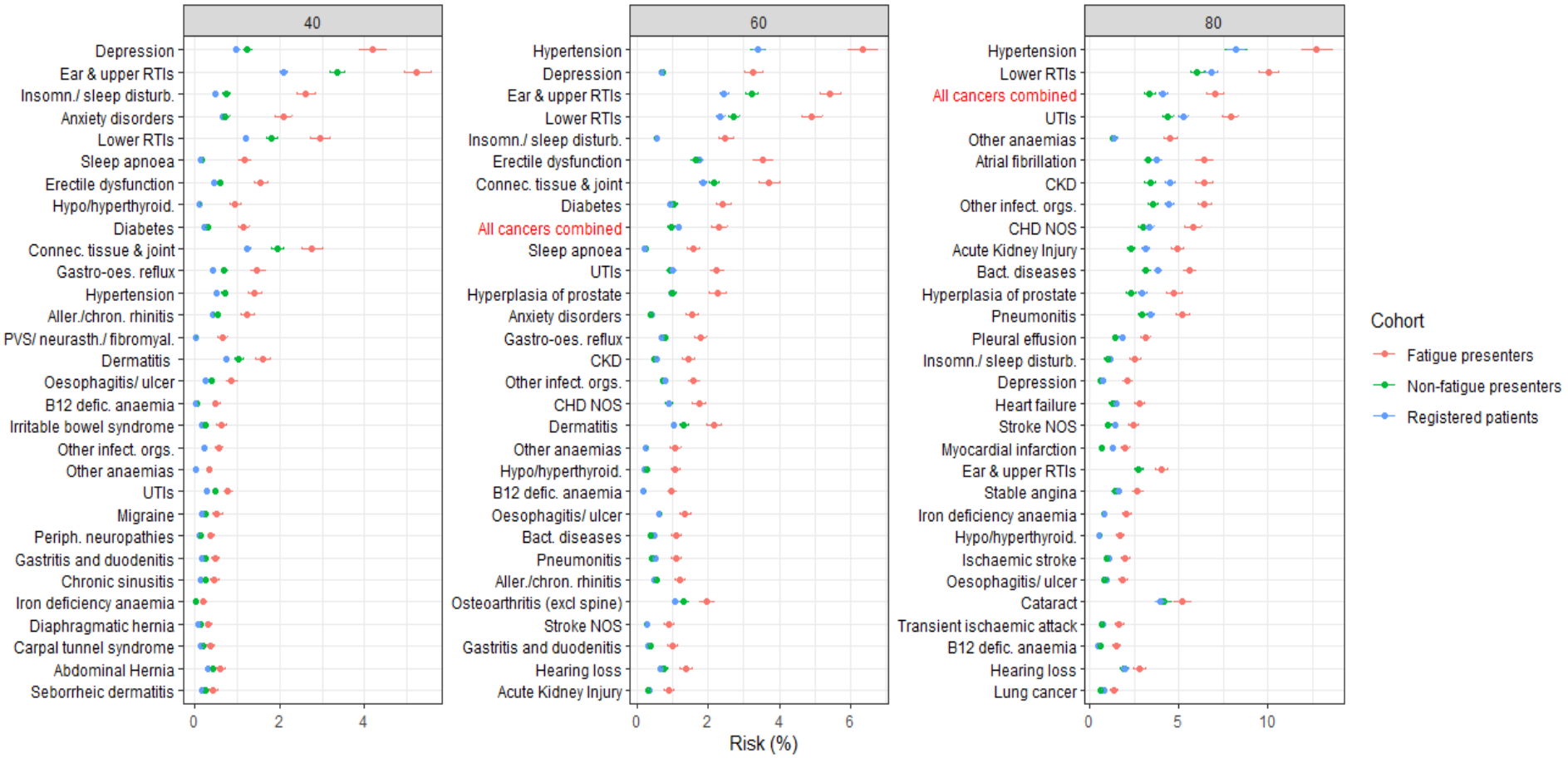



***Figure 6.4. Modelled 1- year risk for diseases in men with fatigue, by cohort (fatigue presenters, non-fatigue presenters, registered patients), for selected ages***

*Top 30 diseases with the greatest absolute excess risk in fatigue presenters compared to non-fatigue presenters in 40, 60, or 80 year olds, ranked by excess risk. All cancers combined are also shown. Excludes patients with a previous diagnosis of each disease.*

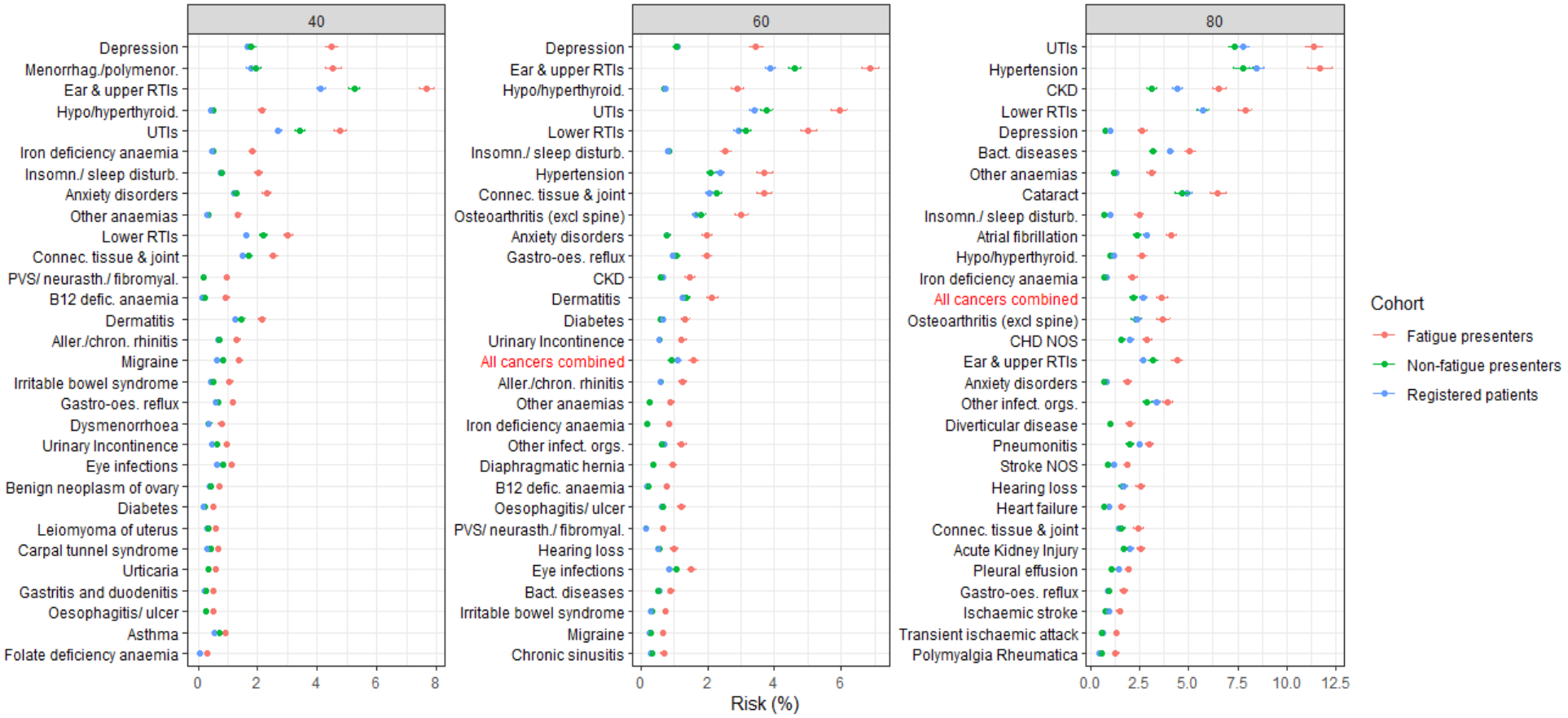



***Figure 6.5. Modelled 1- year risk for diseases in women with fatigue, by cohort (fatigue presenters, non-fatigue presenters, registered patients), for selected ages***

*Top 30 diseases with the greatest absolute excess risk in fatigue presenters compared to non-fatigue presenters in 40, 60, or 80 year olds, ranked by excess risk. All cancers combined are also shown. Excludes patients with a previous diagnosis of each disease.*

### 6.7.5 Supplementary analyses

In this study, I excluded patients with previous diagnoses (in their lifetimes, or for selected infections, in the previous two years). A sensitivity analysis showed that if such patients were included, risk estimates would be higher in fatigue presenters for 209 diseases in men (of which, 140 were statistically significant) and 222 diseases in women (of which, 155 were statistically significant), and additionally would be lower (non-significant) for two diseases (lung and prostate cancer) in fatigue presenters in men. Absolute increases were relatively small (less than 1% higher), except for 17 diseases in men and 17 in women, for which risk was up to 8% higher if including fatigue presenters with previous diagnoses (Appendix 10.6.6).

A supplementary analysis of selected diseases showed that for some diseases (e.g. lower gastro-intestinal (GI) cancer, lung cancer, hypo/hyperthyroidism), risk was concentrated in the first 3 to 6 months after the index date, whereas for others (e.g. insomnia & sleep disturbances, depression, and infections), risk continued to accumulate at a faster rate in fatigue presenters than non-fatigue presenters for the entire 12 months of follow up (Appendix 10.6.9).

## 6.8 Discussion

### 6.8.1 Summary

In this study, I quantify and contextualise estimates of underlying cancer risk in patients with recorded new-onset fatigue in primary care, against a comprehensive map of a wide range of non-neoplastic disease risk. Cancer ranked as a relatively common diagnosis in men with fatigue by 80 years, but it was relatively uncommon in women with fatigue. 100 diseases were more common in fatigue presenters compared to non-fatigue presenters, with short-term risk being substantially higher for (listed in descending order of magnitude in men or women): depression, respiratory tract infections (RTIs), urinary tract infections (UTIs), hypertension, insomnia & sleep disturbances, hypo/hyperthyroidism, and cancer.

### 6.8.2 Strengths and limitations

This study used high quality primary care records from the Clinical Practice Research Datalink (CPRD). This dataset is broadly representative of the UK population regarding age, sex, and ethnicity(145), while linkage to 'gold standard' population-level cancer registration data enabled complete ascertainment of cancer diagnoses for the study cohort(136).

In practice, GPs need to know which diseases to suspect first for individual patients in front of them, i.e. the diseases with the *highest actual risk at specific ages*, regardless of how strong the association is between the disease and fatigue. I have chosen not to rank diseases using this approach, because some of the most common diseases (e.g. cataracts, erectile dysfunction) are not necessarily related to fatigue. Instead, their frequency could reflect their general background risk in primary care presenters. By using a comparison group of 'non-fatigue presenters' to identify diseases with the strongest associations with fatigue, I have chosen an approach to ranking diseases that is most informative for diagnostic guidelines. However, I also report the actual risk of diseases at specific ages in online appendices at https://github.com/rmjlrwh/FatigueRiskMap.

A further strength of this study is the use of two comparator groups to enable stronger inferences within a descriptive study design. Disease risk in fatigue presenters was higher than in non-fatigue presenters. Yet, risk in non-fatigue presenters was not consistently higher than in the wider population of registered patients; it was higher for some diseases but lower for others. While I had hypothesised that a portion of disease risk in fatigue presenters could already be present in patients who consult for any reason (see 'the symptom iceberg' described in Methods(113,114)), the results do not support this hypothesis. Whether general population and non-symptomatic presenter comparator groups can be used interchangeably in future studies requires further consideration. It seems possible that comparisons to general population risk could provide adequate contextual comparison in studies examining disease risk in a symptomatic patient cohort, but this appears to depend on the disease.

I analysed risk of each disease independently from one another, as with similar previous studies(55,96,98). This enabled the quantification of one-year incidence of many difficult to diagnose diseases, which may frequently follow an initial diagnosis of another disease. However, this approach cannot be used to examine whether and how often a disease co-occurs with another, or

how patients progress from one diagnosis to another. Relatedly, elevated risk of a disease following a fatigue presentation does not necessarily imply the disease directly causes fatigue. There are several plausible routes by which we might see elevated risk following a fatigue presentation, including both direct and indirect causal associations, as well as purely incidental findings. Often, the true reason for an association observed among a population of patients diagnosed with a disease, will be a mix of different pathways. In Figure 6.6, I summarise likely causal routes through directed acyclic graphs (see e.g., Greenland et al(200) for a brief introduction to DAGs), using the relationship between fatigue and hypertension as a hypothetical example.

Firstly, it is possible that there is no relationship between fatigue and hypertension (pathway 1). Hypothetically, fatigue might always be caused by another disease such as diabetes, but we would still expect to see a certain number of hypertension diagnoses occurring in fatigue presenters. However, my study design rules out this possibility, as I compared the risk of hypertension in fatigue presenters against non-fatigue presenters and established there is excess risk of hypertension after accounting for age and sex differences. This establishes that there is some kind of association between fatigue presentation and hypertension.

It is possible that hypertension directly causes fatigue in some cases (pathway 2). Alternatively, hypertension could indirectly cause fatigue in other cases, as unmanaged (not previously diagnosed) hypertension can lead to cardiovascular conditions such as heart failure(201) which are potential causes of fatigue (23,37,39,40,53)(pathway 3). Excess risk could also reflect confounding; for example, diabetes is a cause of both hypertension and fatigue(53,201) (pathway 4a). In reality, there are multiple possible pathways (pathway 4b shows one alternative). Independently of whether an association is directly causal, in practice, large excess risk of a particular disease in fatigue presenters could still serve as a potential prompt for clinicians to consider that disease, even if some of the excess risk can be explained through indirect causal pathways. Future studies that aim to fully assess the potential of fatigue to signal the presence of a specific disease should further disentangle the role of confounders. This would require careful assessment of possible causal pathways at the start of the study, and appropriate stratification or adjustment to address potential confounding. One such example could be examining hypertension risk in patients with diabetes (conditional on presence of fatigue), and then considering diagnoses of diabetes and hypertension in patients who presented with fatigue without an existing diabetes diagnosis.

Excess disease risk could also reflect incidental diagnoses discovered through the broad investigations (e.g. ‘tired all the time’ blood tests, blood pressure measurement) that presenting with fatigue could trigger, rather than because these diseases directly cause fatigue (pathway 5). For example, patients may present with fatigue because it is caused by a respiratory tract infection, which is unrelated to hypertension. This may trigger blood pressure measurement to be taken, revealing undetected hypertension which is not responsible for the fatigue. This is possible as prior clinical consensus indicates that hypertension is nearly always asymptomatic(202). It is beyond the scope of this study to identify which diseases with excess risk could be explained by incidental diagnosis, and to what degree.


*Figure 6.6. Hypothetical pathways from hypertension to fatigue presentation*

There is no consensus about the ideal length of follow-up in studies aiming to estimate disease risk after presentation with a given symptom. In my study, due to the large number of disease outcomes studied, I estimated en-bloc risk within 12 months after fatigue presentation, however, excess risk is concentrated in a shorter period (within 3-6 months) for some diseases (e.g. lung cancer, hypo/hyperthyroidism). If a shorter follow up period were used, excess risk of these diseases in fatigue presenters may be even larger, as the contribution of cases due to background incidence unrelated to fatigue will be small (96). In contrast, a longer follow-up will incorporate incident disease unrelated to fatigue presentation occurring in the non-fatigue presenter group, thereby resulting in underestimation of excess risk compared to if a shorter follow up period were used. In addition, the period of excess risk partly reflects how quickly patients are investigated and diagnosed, so en-bloc risk estimates are likely to vary between healthcare systems and over time.

I did not censor patients who died or left CPRD during follow-up. Censoring at death when estimating disease risk is inappropriate, as patients who die are not at risk of developing disease, and such censoring will lead to overestimation of observed disease risk in groups at high mortality risk (e.g. the oldest patients). Various sources of data were available to identify diagnoses in this study (including hospital records and cancer registry), which were not affected by loss to follow up (LTFU) in CPRD. Hence, I did not censor patients if they left their GP practice in CPRD and were LTFU. However, this could lead to underestimation of risk for diseases that are predominantly recorded in primary care data. This is unlikely to be a major problem, as the majority (85%) of fatigue presenters had full one year follow up in CPRD, and 92% had 6 months' follow up.

GPs are more likely to code a symptom, rather than record it as free text, if they deem it to be serious(89). As free text was not available for analysis, some fatigue presenters may not be included in my study, particularly those with milder fatigue. In addition, some diagnoses could have been the result of investigations triggered by another more alarming sign or symptom reported in the same or an earlier consultation. The co-occurrence of an 'alarm' symptom is *a priori* likely to be associated with higher risk of a disease, although a previous study established that co-occurring 'alarm' symptoms for cancer are rarely recorded in fatigue presenters(190). The likely consequence of either of these issues is overestimation of disease risk following a fatigue presentation.

### 6.8.3 Comparison with existing literature

This study, combined with my other recent studies (163,190), substantially enhances the evidence underpinning the UK 2015 National Institute for Health and Care Excellence (NICE) Guidelines for suspected cancer by comparing estimating the risk of all cancer sites combined in patients with fatigue. Previously, estimates were only available for a small number of specific cancer sites, as most of the available studies used case-control designs that identified symptoms that were more frequently recorded before diagnosis in cases (patients diagnosed with a specific cancer), compared to cancer-free matched controls(29).

Previously, it was not known how the risk of cancer compares to other diseases in patients with fatigue, and no existing population-level study has quantified the risk of multiple diagnostic outcomes in patients presenting with fatigue. In a systematic review, Stadje et al. also found that depression and other mental health disorders were more common in fatigue presenters (42). The authors concluded that serious somatic disease was no more common in patients with fatigue than those without, whereas I found cancer risk was substantially higher, especially in older men. The Stadje et al. systematic review reported results for all ages and sexes combined, which has previously been noted as a source of discrepancy in risk studies(55). My finding that older men and women presenting with new-onset fatigue are at elevated cancer risk also concords with previous studies(163,190).

Recent studies have used EHR data to assess the risk of competing (neoplastic and non-neoplastic) diagnostic outcomes in cohorts of patients presenting with other non-specific symptoms such as loss of weight loss and abdominal pain (96,98,164) or other abnormal blood test findings such as thrombocytosis (55). In contrast to the present study, these focussed on specific pre-selected diseases that were deemed serious and related to the symptom. In contrast, my study adapts a comprehensive list of disease phenotypes developed by Kuan et al(170) to track disease incidence and prevalence over the life course, to instead map incident disease risk for a cohort of symptomatic

patients. Regarding specific findings, a previous study found that patients with weight loss were also at increased risk of cancer, depression and hypo/hyperthyroidism (96), but the excess risk of hypertension appears to be unique to fatigue presenters in my study. This might not reflect a stronger causal link between hypertension and fatigue; rather it may reflect a higher propensity among GPs to measure blood pressure in patients with fatigue than those with unexpected weight loss, leading to more incidental diagnoses of hypertension in fatigued patients.

### 6.8.4 Implications for research and practice

My study reiterates previous findings that in older men and women presenting with new-onset fatigue, risk of undiagnosed cancer exceeds 3%, the level of risk at which urgent investigation for suspected cancer in the UK is recommended(16). I show that cancer is a relatively likely diagnosis in older men (aged 80 years) presenting with fatigue, but not in women, in whom the disease spectrum is more dominated by other, non-neoplastic disease. This finding is a combination of several factors, including the lower age-specific risk of cancer in women compared with men. It may also reflect higher prevalence of non-cancerous conditions associated with fatigue in women than men(162), or different fatigue-related help-seeking behaviours(26) resulting in an overrepresentation of men with severe fatigue indicating serious underlying physical disease such as cancer. Therefore, stronger arguments can be made for recommendations to prioritise investigating cancer in older men presenting with fatigue. Whereas in older women with fatigue, doctors could consider safety-netting for cancer or investigating it alongside other possible diagnoses.

This study contextualises cancer risk in patients with new-onset fatigue against a comprehensive risk library of 237 conditions, and so can also inform more general diagnostic guidelines for fatigue. In Table 6.2, I summarise which diseases are most strongly associated with fatigue, additionally flagging which are already included in UK diagnostic guidelines for fatigue (i.e. NICE diagnostic guidelines or BMJ Best Practice guidance). This could enable guidelines to go beyond the simple listing of differential diagnoses, by attaching a quantitative ranking of the risk of different diseases. For example, given my findings, the highest ranks could relate to depression, lower and upper RTIs, insomnia & sleep disturbances, hypo/hyperthyroidism (women), and cancer (men).

***Table 6.2. Excess risk of cancer and diseases with >1% absolute excess risk (AER) in male or female fatigue presenters compared to non-fatigue presenters, ranked by AER***

**Adjusting for age differences between fatigue presenters and non-fatigue presenters. **Diseases included in current UK diagnostic guidance for fatigue published in either NICE or BMJ* (37,54). **** Cancer is highlighted for women despite not having an AER > 1%.*

| Disease | Overall absolute risk in FPs (%) | Overall absolute risk* in NFPs (%) | Overall AER* in FPs vs NFPs (%) | Rank of overall AER* | Included in current guidelines** |
|---|---|---|---|---|---|
| **Men** | | | | | |
| Depression | 3.21 | 0.83 | 2.38 | 1 | Y |
| Lower Respiratory Tract Infections | 5.55 | 3.26 | 2.28 | 2 | Y |
| Hypertension | 5.09 | 2.91 | 2.18 | 3 | N |
| Insomnia & sleep disturbances | 2.55 | 0.7 | 1.85 | 4 | Y |
| Ear and Upper Respiratory Tract Infections | 4.96 | 3.16 | 1.79 | 5 | Y |
| Urinary Tract Infections | 3.21 | 1.71 | 1.49 | 6 | N |
| **All cancers combined** | **2.59** | **1.16** | **1.43** | **7** | **Y** |
| Other or unspecified infectious organisms | 2.56 | 1.3 | 1.26 | 8 | Y |
| Other anaemias | 1.68 | 0.43 | 1.25 | 9 | Y |
| Erectile dysfunction | 2.12 | 0.93 | 1.19 | 10 | N |
| Chronic kidney disease | 2.12 | 0.98 | 1.14 | 11 | Y |
| Anxiety disorders | 1.53 | 0.48 | 1.05 | 12 | Y |
| Diabetes | 1.82 | 0.78 | 1.04 | 13 | Y |
| Connective & soft tissue disorders | 2.9 | 1.88 | 1.02 | 14 | Y |
| **Women** | | | | | |
| Depression | 3.64 | 1.28 | 2.36 | 1 | Y |
| Urinary Tract Infections | 6.66 | 4.42 | 2.24 | 2 | N |
| Ear and Upper Respiratory Tract Infections | 6.64 | 4.54 | 2.1 | 3 | Y |
| Hypo or hyperthyroidism | 2.43 | 0.67 | 1.76 | 4 | Y |
| Lower Respiratory Tract Infections | 4.79 | 3.28 | 1.52 | 5 | Y |
| Insomnia & sleep disturbances | 2.25 | 0.8 | 1.46 | 6 | Y |
| Anxiety disorders | 2.08 | 0.97 | 1.11 | 7 | Y |
| Hypertension | 3.06 | 1.96 | 1.1 | 8 | N |
| Iron deficiency anaemia | 1.48 | 0.42 | 1.07 | 9 | Y |
| Other anaemias | 1.52 | 0.48 | 1.05 | 10 | Y |
| *Next 10 diseases with < 1% AER not shown* | - | - | - | - | - |
| **All cancers combined***** | **1.42** | **0.9** | **0.52** | 21 | Y |

I also found over 1% excess risk of other diseases which are not listed in existing guidelines, including urinary tract infections, hypertension, erectile dysfunction, connective & soft tissue disorders, and cataracts. No previous evidence I reviewed identified these as underlying causes of fatigue(21–24,37,39–42,53,186). Further research is needed to elucidate possible mechanisms explaining such excess risk, including: a) direct causal pathways, b) indirect causal pathways, c) confounders such as comorbidities, and d) incidental diagnoses. If pathways a-c dominate for a particular disease, fatigue could still serve as a potential prompt for clinicians to consider that disease, regardless of whether the underlying mechanism can be explained. For diseases where it is suspected that incidental diagnosis plays a substantial role in the observed excess risk, it cannot be recommended that patients be investigated for the disease on the basis of fatigue presentation alone. As it was beyond the scope of this study to assess the role of these causal mechanisms, further clinical input and research is needed before adding any of these diseases to diagnostic guidelines for fatigue.

In addition, the findings highlight that patients with fatigue who are at greater than 3% risk of cancer (the level of risk at which urgent investigation for suspected cancer in the UK is recommended(16)) may also be at notable absolute risk of other conditions that typically require urgent hospital referral for diagnosis and/ or management, hence follow-up diagnostic strategies are important, particularly when cancer has been excluded. For example, an 80 year old man presenting with fatigue has a 7% risk of cancer, 5% of pneumonitis, and 5% of acute kidney injury. Urgent referral for further investigation through non-specific referral pathways (e.g. Rapid Diagnostic Centres in the UK) may have considerable value for these patients. The findings highlight the importance of expanding multidisciplinary diagnostic services for patients with non-specific symptoms. Conversely, while the risks of various consequential diseases such as depression is high in younger men (aged 40 years) with fatigue, risks of potentially urgent diagnoses are low (< 1%).

Given the elevated risk of a wide range of diseases in patients with new-onset fatigue, strong conclusions about the most likely working diagnosis cannot be made easily using the mere presence of fatigue combined with the patient's age and sex. In practice, GPs need to consider other clues from the patient's medical history, presenting features or additional tests(37,53,54). When serious disease is suspected, GP access to direct tests(203) and referral to Rapid Diagnostic Centres (RDCs)(46,47) are options that could help healthcare providers quickly narrow down the most appropriate referral pathways, and if considering cancer, the likely primary site.

## 6.9 Chapter summary

For the first time, my study quantifies and contextualises short term cancer risk in patients presenting to their GP with new-onset fatigue, relative to the risk of other possible diagnoses. Strong arguments can be made for recommendations to prioritise investigating cancer in older men presenting with fatigue, for whom cancer is a relatively common diagnosis. In contrast, in older women with fatigue, in the absence of other signs and symptoms of cancer, doctors could consider safety-netting for cancer or investigating it alongside other possible diagnoses. Diagnostic investigation and referral guidelines for new-onset fatigue could be informed by the risks reported by this study, which found that diseases such as depression, lower and upper respiratory tract infections (RTIs), insomnia & sleep disturbances, hypo/hyperthyroidism (women), and cancer (men) were most strongly associated with fatigue.

# 7. Chapter 7: Discussion

## 7.1 Chapter summary

In this chapter, I bring together the findings across my primary empirical studies on fatigue. I discuss the studies' strengths and weaknesses, compare results to existing literature, and discuss implications for clinical practice and future research. The literature review of diagnostic window studies, which helped to focus my research by revealing the potential to diagnose cancers earlier through primary care, is not discussed in this chapter but in Chapter 2.

## 7.2 Key findings

### 7.2.1 Risk of cancer in patients with fatigue

In Chapter 4, I established that without considering accompanying symptoms, new-onset fatigue was associated with one-year cancer risk exceeding 3% in older men and women, which is the current threshold for urgent specialist referral in the UK (although the risk of any single cancer site was substantially smaller). The risk was significantly higher than the background risk in the general population in patients of the same age and sex. It also revealed that the period of excess risk after the first fatigue presentation was limited to up to nine months (with the majority of risk concentrated in the first three months, after which it returned to the background risk in the general population.

### 7.2.2 Risk of cancer in patients with fatigue and other vague symptoms

In Chapter 5, I restricted the study cohort to patients presenting with fatigue without accompanying alarm symptoms or anaemia, and followed up patients for nine months (the period of excess risk established in the first study). Cancer risk did not exceed 3% when examining only those fatigued patients without alarm symptoms or anaemia, except in men aged 73 years and over.

However, when older patients presented with fatigue in combination with another vague symptom, risk exceeded the 3% urgent referral thresholds. In men, risk exceeded 3% in patients aged 70 years and over with any of the 19 vague symptoms studied, and was highest in those with fatigue combined with either weight loss (exceeding 3% from 59 years), abdominal pain (from 65 years), constipation (from 67 years) or other upper gastrointestinal symptoms (from 67 years). In women, risk exceeded 3% only when fatigue was combined with weight loss (from 65 years), abdominal pain (from 79 years), or abdominal bloating (from 80 years).

### 7.2.3 Risk of specific cancer sites and other diseases in fatigued patients

In Chapter 4, I established that fatigue presentation was not strongly predictive of any single cancer, although certain cancers (e.g. leukaemia, pancreatic and brain cancer) were generally over-represented compared to their incidence in the general population.

In Chapter 6, cancer was relatively uncommon in female fatigue presenters compared to other (non-neoplastic) diagnoses, whereas it was one of the seven most common diagnoses in men with fatigue. The excess risk of cancer increased with age; by 80 years, in men it was the disease with the 3rd greatest excess risk compared to non-fatigue presenters (with 7.0% absolute risk in fatigue presenters), and in women it was the 13$^{th}$ (3.6% absolute risk). Men and women aged 80 years with fatigue were also at notably high risk (ranging from 1 to 5%) of other diagnoses, including pneumonitis, acute kidney injury, stroke, chronic kidney disease, and coronary heart disease.

The absolute excess risk of 100 out of 237 diseases studied was higher in fatigue presenters than non-fatigue presenters. These included (listed in descending order of magnitude in men or women):

depression, respiratory tract infections (RTIs), urinary tract infections (UTIs), hypertension, insomnia & sleep disturbances, hypo/hyperthyroidism, cancer, chronic kidney disease. When ranked by their frequency at specific ages, depression, insomnia & sleep disturbances, and hypo/hyperthyroidism (women only) were relatively common in fatigue presenters compared to non-fatigue presenters.

## 7.3 Strengths and limitations

The methodological issues involved in conducting cohort studies of disease risk using electronic health records (EHRs), and how I addressed them, are discussed in detail in Chapter 3. In this section, I summarise the strengths and limitations of the methods, their potential impact on results, and discuss possible solutions in future similar studies.

### *7.3.1* Strengths and limitations of CPRD

A strength of my studies is the use of high-quality electronic health records from Clinical Practice Research Datalink (CPRD) GOLD, which are broadly representative of the age, sex, and ethnicity distribution of the UK population. However, since 2013, the number of practices participating in CPRD GOLD has fallen which has diminished the sample size of my studies in the latter study years (145,146). Future research could externally validate the research by examining whether the risk estimates reported in my studies also hold in CPRD Aurum, which is based on a GP software system (EMIS WEB®) that is more widely used in recent eras, and also in UK databases based on other GP software systems such as OpenSAFELY-TPP (https://www.opensafely.org/about/).Linkage to population-level cancer registration (CR) data offered 'gold standard' ascertainment of cancer diagnoses, even after patients were lost to follow up within CPRD. When studying multiple disease outcomes, Hospital Episodes Statistics (HES) data also offered continued follow up of conditions which are often diagnosed and/ or treated in secondary care (e.g. heart failure)(204). Nonetheless, diagnoses usually recorded in primary care (e.g. depression, diabetes) could potentially be undercounted due to loss to follow up in CPRD, although in practice this is unlikely to be a major problem, as the majority (85%) of fatigue presenters had full one year follow up in CPRD, and 92% had 6 months' follow up.

Some instances of a patient's presentation with fatigue may not have been coded by the GP in the patient's record. GPs are more likely to record alarm symptoms as coded entries rather than free text (and free text data are not available to researchers in the UK due to information governance constraints) when there is a suspicion for cancer (89). It is not known whether fatigue is subject to the same recording patterns, although it is reassuring to note that for another non-specific symptom (abdominal pain) no systematic under-recording was found (89). There is no existing UK study that systematically compares cancer risk (for all cancers combined, and by cancer site) in cohorts of patients with non-specific symptoms whose symptoms were recorded in coded data versus those recorded only in free text. As access to free text records for research is generally restricted in the UK, such a study would be extremely valuable.

The external validity of CPRD could vary over time according not only to doctor's recording practices, but to patient presentation behaviour and changes in the true underlying prevalence of different underlying causes in the population. For example, my studies cover a period before the SARS-CoV-2 pandemic began in 2020. Given the prevalence (and likely heightened public awareness) of fatigue both as an acute symptom of the disease and the defining feature of 'long-covid' syndrome(205), the association between fatigue and short-term cancer risk could have been weakened during the pandemic and for a time afterwards.

### 7.3.2 Comparisons

As discussed in Chapter 3, the choice of comparisons and statistical methods in primary care electronic health record (EHR) based studies of short-term disease risk in cohorts of patients presenting with symptoms should be based in the study purpose, which can be categorised (with some overlap) into three broad themes; descriptive risk studies, diagnostic value studies, and risk prediction model studies. I used comparison groups to contextualise risk and enable stronger inferences within a descriptive design; hence I compared disease risk in fatigue presenters against the background risk in the general population (or registered patients) and also to patients presenting without fatigue. This represents a novel comparison approach, which indicated that a) fatigue had diagnostic value over and above simply presenting to primary care for any other reason, and b) presenting to primary care is not necessarily associated with increased disease risk (since risk in non-fatigue presenters was higher than in the wider population of registered patients for some diseases but lower for others). This suggests that the size and shape of the 'symptom iceberg' described in Section 3.5.2 depends on the disease, and that for some diseases, general population risk could provide adequate contextual comparison in studies examining disease risk in a symptomatic patient cohort (as used in Chapter 4 and by Price et al(48)).

The comparisons I used were descriptive rather than explanatory, so the links between primary care presentation and increased disease risk, and between fatigue presentation and increased disease risk are not necessarily causal and could be caused by confounding (e.g. by comorbidity) or incidental diagnoses. I aimed to simply establish whether fatigue could be a possible sign or symptom of disease, so have controlled only for age and sex when comparing fatigue presenters and non-fatigue presenters. However, as discussed in Chapter 3, aetiological studies that aim to quantify the 'diagnostic value' of a symptom (i.e. more confidently assert whether it is linked to an underlying disease independently of the role of other associations) ought to control for additional factors (e.g. comorbidities) (e.g. Nicholson et al (97)).

### *7.3.3* Statistical methods

I estimated crude cumulative incidence, that is, risk as a proportion of patients in the denominator at the start of follow up, regardless of whether they were later lost to follow up (LTFU) in CPRD. This approach ensured that the study cohort represented all patients presenting with fatigue in primary care, rather than excluding those who moved practice (noting that patients who move practice are potentially less healthy on average than those with continuous single practice enrolment (116)). This approach was suitable as full ascertainment of cancers was possible using linked national cancer registry data, and the effects of LTFU on ascertainment of other (non-cancer) disease diagnoses were somewhat mitigated by the inclusion of secondary care (Hospital Episodes Statistics Admitted Patient Care (HES APC)) data. However, the risk of diseases that are predominantly diagnosed in primary care (e.g. depression) could still be underestimated in patients who were LTFU in CPRD. Such underestimates could be exacerbated in groups of patients who change GP practice more often than average.

I also did not censor patients who died during follow up. This generated risk calculations that were easily interpretable for GPs and reflect a patient's 'real' risk shortly after an initial consultation(115). This means that although disease risks estimated in my thesis are more relevant to clinical practice,

they should be interpreted with caution in patients at high risk of death (e.g. aged 90 years and over). In these groups, disease risk may be low because death commonly occurs first. However, as the focus of the studies were to describe actual disease risk in comparison to referral thresholds, it would not be appropriate to censor patients who died.

### 7.3.4 Fatigue cohort definition

To produce risk estimates relevant to GPs, I aimed to ensure the study population broadly represented patients attending primary care with new-onset fatigue, minimising the likelihood that it was attributable to a recently diagnosed condition or disease (including cancer) or its treatment. Therefore, in Chapter 4, I included patients only if they had an eligible fatigue record that occurred more than a year following another fatigue record or cancer diagnosis.

In Chapter 6, I also aimed to minimise the possibility that estimates of disease risk were influenced by a prior diagnosis, which was particularly important for chronic diseases such as Inflammatory Bowel Disease (IBD) or HIV. For each of the 237 diseases studied (say 'disease x'), I defined a 'disease x-free' cohort of patients without a prior diagnosis of that specific disease ('x'), and estimated risk in that cohort. For most diseases, including cancer diagnoses, I excluded patients with a prior diagnosis at any point before their index date (with the exception of some common infections, for which a two-year lookback was used). A sensitivity analysis showed that including patients with a prior diagnosis would result in statistically significant higher disease estimates compared with those observed in the main analysis for 140 diseases in men and 155 in women. Future research would benefit from more information about how subsequent risk varies according to the exact preceding time period used to define 'disease-free' cohorts, and how this varies by disease.

In order to adequately identify patients with previous diagnoses so that they could be excluded, all patients included in the study are required to have adequate follow up before index date. Therefore, only patients who were registered to their practice for at least one year (Chapters 4 & 5) or two years (Chapter 6) before their index date could be included in the study. This could lead to the selection of a healthier population at lower mortality risk(116).

### *7.3.5* Co-occurring feature definition

A strength of Chapter 5 is that I identified fatigue presenters with 'co-occurring' symptoms recorded 3 months before to one month after the index fatigue presentation, following a thorough theoretical and practical exploration of the appropriate time period to include. Including symptoms after an index date can introduce immortal time bias, where patients are required to have 'survived' free of the outcome for that period to time to be included in the symptom group, which can bias risk estimates(131). My inclusion of a short time period following the index date minimised such bias, while incorporating a short lag to capture the delayed recording of symptoms that may be present in the patient's first consultation about fatigue. Meanwhile, I demonstrated for the first time that cancer risk estimates would be lower if using longer look-back periods for including co-occurring symptoms, as symptoms further in the past are less likely to be related to the underlying cause that prompts a presentation with new-onset fatigue It was deemed that a three-month lookback period offered a good balance between this consideration and the need to generate adequate cohort sizes of patients with each co-occurring symptom.

In general, for studies such as mine which primarily aim to describe risk in a symptomatic cohort and provide foundational risk stratification to inform diagnostic guideline recommendations (see Section 3.5.1), the simple categorisation of patients into co-occurring symptom groups is appropriate. In studies to date, this meant an inclusion time window was chosen to look before and after the index presentation. There is potential to refine these methods in future descriptive studies, for example by

giving patients a ‘co-occurrence likelihood score’ based on how closely the record(s) of additional symptoms occurred in relation to the index symptom (as opposed to using a binary variable flagging whether they did or not have the co-occurring symptom).

### *7.3.6* Outcome definition

To generate findings that were easily interpretable for clinicians and guideline policy makers, I calculated ‘en bloc’ estimates of disease risk within a specified period of time (e.g. 12 months in the first and third study) following the index date. Unlike most similar studies (except Nicholson et al. and Withrow et al.(96,97)), I identified the period of excess cancer risk following the first fatigue presentation, which provided evidence for the nine month follow up period used in Chapter 5.

A limitation of using an ‘en bloc’ follow up period is that disease risk is sensitive to the time period chosen(96). In Chapter 6, I showed that excess risk was concentrated in a fairly short period (within 3-6 months) for some diseases (e.g. lung cancer, hypo/hyperthyroidism). For these diseases, using a longer 12-month follow up period likely underestimated the excess risk in fatigue presenters relative to their background disease risk, particularly their shorter-term excess risk.

### 7.3.7 Phenotype development

The process I used to quality assure symptom and disease phenotypes is another strength of the studies, and ensured that the code lists used remained up to date and sufficiently sensitive.

When reviewing the fatigue phenotype in Chapter 4, I found code list to be stable over time and sufficiently sensitive, missing only one relatively uncommon code that was subsequently added in the following studies.

In chapter 5, I maximised the sensitivity of phenotypes for ‘co-occurring’ symptoms by combining existing code lists for the same symptom from multiple sources. To my knowledge, the 35 symptom phenotypes I used represent the largest collection of symptoms used in a single EHR-based cohort study to date. However, when excluding patients with co-occurring ‘alarm’ symptoms, it is possible that a small minority of patients could be erroneously included if they had one of 12 alarm symptoms for which Read code lists were unavailable (Appendix 10.5.3), though these symptoms are unlikely to be common in practice. The 12 missing alarm symptoms comprised the following: vulval bleeding, prostate feels malignant on examination, anal mass or anal ulceration on examination, vaginal mass, vulval lump or ulceration, appearance of cervix consistent with cervical cancer, lip or oral cavity lump, oral cavity red or red/ white patch erythroplakia or erythroleukoplakia, alcohol induced lymph node pain, skin lesion, penile mass, other penile symptoms affecting the foreskin or glans.

Finally, to identify relevant disease outcomes, I combined phenotypes published through large scale phenotyping projects with other existing code lists for diseases potentially related to fatigue. I updated these code lists and maximised their sensitivity by searching for common Read V2 codes, International Classification of Diseases 10th Revision (ICD 10) codes, and Office of Population Censuses and Surveys Classification of Interventions and Procedures version 4 (OPCS4) codes occurring in any of the datasets.

## 7.4 Comparison with existing literature

### 7.4.1 Risk of cancer in patients with fatigue

Available evidence underpinning current National Institute for Health and Care Excellence (NICE) guidelines has so far only examined the positive predictive value (PPV) of fatigue for diagnosis of a small number of specific cancer sites(16). My studies substantially enhance previous evidence regarding the risk of present but as-yet-undetected cancer among patients presenting to primary care with fatigue, as it is the first to examine risk of cancer overall.

According to a systematic review, previous studies (generally using case-control as opposed to cohort study designs) found that fatigue was associated with specific cancers such as leukaemia, lung and kidney cancers(206). However, a widely-used risk prediction tool (QCancer) reported that fatigue was not a significant independent predictor of cancer within 24-months, unlike other non-site specific symptoms and features, such as weight loss, appetite loss, and venous thromboembolism(44,45). Differences to the findings of my research could arise from various factors, including differences in the data source, length of follow-up, and adjustment for other presenting symptoms or other variables.

### *7.4.2* Risk of cancer in patients with fatigue and other vague symptoms

To my knowledge, this research is the first to examine cancer risk in a cohort of patients presenting with a non-specific symptom, in the absence of any recorded alarm symptoms for cancer. It is also the first to characterise the co-occurrence of other vague symptoms in patients presenting with fatigue, and the associated cancer risk. Together with other recent evidence(44,45,48,49,96,163), the findings establish abdominal pain, weight loss, and fatigue as vague symptoms that confer a substantial risk of cancer (though this comprises smaller risks for specific cancer sites) often exceeding risk thresholds for further cancer investigation, particularly when patients presenting in combination with other vague symptoms.

In addition, older males without alarm symptoms or anaemia and with fatigue–constipation or fatigue–other upper GI symptoms (which included dyspepsia, nausea, vomiting, haematemesis, loss of appetite) and older females with fatigue–abdominal bloating were also at elevated (>3%) risk of cancer. These findings are concordant with prior literature examining some of these abdominal symptoms either alone(44,45) or in combination with weight loss(49) or abdominal pain(48).

### 7.4.3 Risk of specific cancer sites and other diseases in fatigued patients

My thesis substantially enhances existing evidence underpinning current NICE Guidelines for suspected cancer by comparing risk of the full range of possible cancer sites in patients with fatigue. Previously, estimates were only available for a small number of specific cancer sites, as most of the available studies used case-control designs that identified symptoms that were more frequently recorded before diagnosis in cases (patients diagnosed with a specific cancer), compared to cancer-free matched controls(29).

Previously, it was not known how the risk of cancer compares to other diseases in patients with fatigue, and no existing population-level study has quantified the risk of multiple diagnostic outcomes in patients presenting with fatigue. Recent studies have used electronic health record (EHR) data to assess the risk of competing (neoplastic and non-neoplastic) diagnostic outcomes in cohorts of patients presenting with other non-specific symptoms such as loss of weight and abdominal pain (96,98,164) or other abnormal blood test findings such as thrombocytosis (55). In contrast to the present study, these focussed on specific pre-selected diseases that were deemed serious and related to the symptom. My research is the first to comprehensively map the risk of incident disease for a cohort of symptomatic patients, and the first population-level study to quantify the risk of multiple diagnostic outcomes in patients presenting with fatigue.

Many of the conditions that I identified as having the greatest excess risk in fatigued patients are already included as potential diagnoses in UK diagnostic care guidance for fatigue(54), including cancer, depression, RTIs, UTIs, hypertension, insomnia & sleep disturbances, hypo/hyperthyroidism, cancer, and chronic kidney disease (54). Current guidance lists potential diagnoses, but cannot yet quantify which diagnoses are most likely and should be considered first. My research enables the development of such guidelines (as well as diagnostic strategy algorithms and future health economics analyses) by quantifying the risk of diseases and ranking them by absolute and excess risk, by patient age and sex.

I also found over 1% excess risk of other diseases which are not listed in existing guidelines, including urinary tract infections, hypertension, erectile dysfunction, connective & soft tissue disorders, and cataracts. No previous evidence I reviewed identified these as underlying causes of fatigue(21–24,37,39–42,53,186). As discussed in Chapter 6, excess risk of a disease could reflect different mechanisms including: a) direct causal pathways, b) indirect causal pathways, c) confounders such as comorbidities, and d) incidental diagnoses. For diseases where it is suspected that incidental diagnosis plays a substantial role in the observed excess risk, it cannot be recommended that patients be investigated for the disease on the basis of fatigue presentation alone.

## 7.5 Implications for policy and practice

### 7.5.1 Recommendations for UK diagnostic guidelines for suspected cancer

My research shows that overall cancer risk exceeds current National Institute for Health and Care Excellence (NICE) thresholds (> 3%) for urgent investigation for suspected cancer in the UK(16) in older men and women presenting with new-onset fatigue, but when considering only those without alarm symptoms or anaemia, it does not (except for men aged 73 and over). Nevertheless, in these patients, when fatigue is recorded alongside other vague symptoms, risk does exceed referral thresholds, ranging from ages 59-80 years depending on the symptom and patient's sex; in particular, fatigue combined with weight loss, abdominal pain, abdominal bloating (women), constipation (men), or other upper gastrointestinal symptoms (men). These groups could be considered for inclusion in NICE referral guidelines for suspected cancer.

Fatigue presentation alone is not strongly predictive of cancer of any single organ; even at ages where overall cancer risk exceeds 3%, this comprised much lower risks of any single cancer site. In the UK, urgent 'two-week-wait' referral routes are cancer site-specific, being grouped within 16 medical specialties (e.g. haematology, lower GI etc.) (188). This means that, in practice, urgent referrals to one of the specialties need to be guided by other signs and symptoms, and the GPs' clinical assessment of which cancer is most likely among fatigue presenters with suspected underlying cancer. Future analysis of site-specific cancer risk in fatigue presenters incorporating other signs and symptoms of disease might produce more refined evidence to support such referral decisions.

These findings emphasise the need to establish non-specific diagnostic care pathways for patients with fatigue or other non-specific symptoms. Following the introduction of large-scale service innovations aimed at achieving fast diagnostic resolution in patients with non-specific symptoms in Denmark, the NHS has introduced Rapid Diagnostic Centres (46). However, these may be limited by capacity constraints. Currently (June 2023), there is no publicly available information about how many patients with a suspected cancer diagnosis are received by Rapid Diagnostic Centres (RDCs), though I am aware through personal communication that many NHS Cancer Alliance regions do not yet have a non-specific pathway service.

The findings regarding the risk of cancer compared to other diagnoses adds important context when considering additions to NICE referral guidelines. Cancer is a relatively likely diagnosis in older men (aged 80 years) presenting with fatigue, but not in women, where the disease spectrum is more dominated by other, non-neoplastic disease, compared with men. This finding is a combination of several factors, including the lower age-specific risk of cancer in women compared with men. It may also reflect higher prevalence of non-cancerous conditions associated with fatigue in women than men(162), or different fatigue-related help-seeking behaviours(26) resulting in an overrepresentation of men with severe fatigue indicating serious underlying physical disease such as cancer. Therefore, stronger arguments can be made for recommendations to prioritise investigating cancer in older men presenting with fatigue alongside other vague symptoms. Whereas in older women with fatigue combined with other vague symptoms, doctors could consider safety-netting for cancer or investigating it alongside other possible diagnoses.

The benefits of ruling out serious physical disease such as cancer must also be weighed against the risks of over investigation in older patients with non-specific symptoms. For instance, an 80-year old

man presenting with fatigue and no alarm symptoms has over 3% risk of being diagnosed with cancer in the next year. Yet, this could largely reflect his background risk (unrelated to new onset fatigue), as one-year cancer risk for men in the general population aged over 80 years already exceeds 3% (Chapter 4). Applying a 3% referral threshold without consideration of a patient's baseline risk could lead to frequent, unwanted investigations, leading to negative mental and physical impacts for patients and increased healthcare service costs(19,186). Referral decisions need to weigh these benefits and risks, incorporating patient preferences supported by accurate communication of diagnostic uncertainty(187). Further thought is also needed about whether urgent referral should be recommended if the risk of all cancers combined exceeds 3%, and its potential implications for over investigation, since the 3% threshold has so far generally applied to the risk of individual cancer sites in NICE referral guidelines (16).

### 7.5.2 Recommendations for UK RDC and NICE fatigue guidelines

In the UK, the standard referral route for urgent investigation for suspected cancer is the two-week-wait (TWW) pathway, which requires a single cancer site to be suspected. More recently, Rapid Diagnostic Centres (RDCs) have been introduced, which are one-stop diagnostic centres that aim to achieve diagnostic resolution in patients with non-specific symptoms such as fatigue(46). My research shows that although certain cancers (e.g. leukaemia, pancreatic and brain cancers) are generally over-represented in fatigue presenters, the most common cancers in the general population still accounted for a large proportion of cancers diagnosed in fatigue presenters. Since fatigue presentation alone is not strongly predictive of cancer of any single organ, investigation through non-cancer-site specific routes such as RDCs could be appropriate when cancer is suspected and there are no accompanying site-specific symptoms. However, the deployment of RDCs may be constrained by capacity; the numbers of patients currently referred via this pathway is not yet publicly available, and many parts of England do not yet have an RDC (June 2023).

Moreover, some groups of patients with fatigue could benefit from referral to an RDC because of their heightened risk of multiple diseases that usually require secondary care referral for diagnosis. For example, an 80 year old man presenting with fatigue has approximately 7% risk of cancer, 5% of pneumonitis, and 5% of acute kidney injury (Chapter 6.8.4, Appendix 10.6.8). In such cases, where there is a need for investigation in secondary care but the initial working diagnosis is unclear, patients could be referred to an RDC, subject to other signs, symptoms, and the results of primary care diagnostic tests.

Findings from my thesis can also inform general UK diagnostic care guidance for fatigue(54), which support GPs' referral decisions both to RDCs and other routes. In Chapter 6 (Table 6.2), I summarise which diseases are most strongly associated with fatigue, and additionally flag which are already included in UK diagnostic guidelines for fatigue (i.e. NICE diagnostic guidelines or BMJ Best Practice guidance). As shown in Box 1, current BMJ Best Practice guidance lists a broad range of differential diagnoses, which are not currently ranked. The results of my study could enable guidelines to go beyond the simple listing of differential diagnoses, by attaching a quantitative ranking of the risk of different diseases. For example, given my findings, the highest ranks could relate to depression, lower and upper RTIs, insomnia & sleep disturbances, hypo/hyperthyroidism (women), and cancer (men).

***Box 1. Aetiology of fatigue, adapted from BMJ Best Practice Guidance***(37)

**Cancer**

**Cardiovascular disease:** Heart failure, acute myocardial infarction, atrial fibrillation

**Drugs and toxins:** Recreational drugs, antihistamines, antihypertensives, anti-arrhythmics, antidepressants, anti-emetics, antiepileptics, corticosteroids, diuretics, and neuroleptic agents, ticagrelor, chronic alcohol misuse, heavy metal toxicity

**Endocrine disorders**: Hypothyroidism, diabetes mellitus, Addison's disease, vitamin D deficiency, hypopituitarism, acromegaly, growth hormone deficiency, hyperthyroidism, Cushing's syndrome, diabetes insipidus

**Gastrointestinal disorders:** Coeliac disease, chronic liver disease, inflammatory bowel disease, irritable bowel syndrome

**Haematological disorders:** Anaemia, chronic myeloid leukaemia, myelodysplastic syndrome, lymphoma, heavy metal toxicity

**Idiopathic causes:** Chronic fatigue syndrome, systemic exertion intolerance disease

**Infectious disease:** Epstein-Barr virus (EBV), HIV, COVID-19, Lyme disease, cytomegalovirus, toxoplasmosis, Q fever, brucellosis, tuberculosis, coxsackie B virus, chlamydia, mycoplasma, influenza virus

**Neurological disorders:** Parkinson's disease, stroke, multiple sclerosis, lateral amyotrophic sclerosis, myasthenia gravis, dystonias, myopathies

**Psychiatric and psychosocial disorders:** Depression, anxiety and somatisation disorders

**Pulmonary disease:** COPD, sarcoidosis, asthma, pulmonary HTN, pleural disease, and pneumonitis

psychosocial stressors

**Renal disorders:** Haemodialysis, renal failure

**Rheumatological disorders:** Systemic lupus erythematosus, fibromyalgia, rheumatoid arthritis

**Sleep disorders:** Insomnia, obstructive sleep apnoea/ hypopnoea syndrome, obesity hypoventilation syndrome, restless legs syndrome

### 7.5.3 Recommendations for future research

*Risk stratification to support referral guidelines for cancer*

Although my research confirms that fatigue is associated with elevated risk of cancer and a wide range of other diseases, there are a number of areas where further research is needed to further support the development of diagnostic guidelines for cancer.

Considered in isolation, the potential to use fatigue to distinguish between the most likely diagnoses is poor. In practice, GPs need to consider other clues from the patient's medical history, presenting features or the results of additional tests(37,53,54). When serious disease is suspected, GP access to direct tests(203) are options that could help them quickly narrow down the most appropriate referral pathways, and if considering cancer, the likely primary site. Recent development of Multi-Cancer Early Detection (MCED) tests holds particular promise for diagnosis in patients presenting only with vague symptoms such as fatigue, as some are able to indicate the likely primary cancer site. Research will be needed to judge their additive predictive value in such clinical scenarios, over and above information that can be gleaned by existing commonly-used blood tests. Future research should assess their potential effectiveness in patients with vague symptoms, compared to currently available tests and information about age- and sex-specific risk following symptomatic presentation. Their broader health services implications should be assessed, such as their potential to increase or reduce subsequent need for other investigations and potential over-testing in patients with vague symptoms.

My research has demonstrated that electronic health records (EHRs) can be used to examine a wide range of multiple diseases and cancer sites associated with a single symptom, as well as how risk of a single disease (cancer) varies according to a range of co-occurring symptoms in combination with fatigue. EHRs could be further harnessed to compare the risk of different diagnostic outcomes in the presence of multiple diagnostic features in a single study, to assess whether additional features add discriminatory power between potential diagnoses in patients with fatigue. In addition, future research could characterise disease risk in patients with new-onset fatigue by combining other co-occurring features, such as blood test results or comorbidities. Such evidence is paradigmatically important as it could be transferred to patients presenting with other vague symptoms (e.g. weight loss, abdominal symptoms, shortness of breath, etc.).

While contributions from my research and other authors have greatly increased available evidence on non-specific symptoms that can support future updates to NICE referral guidelines for suspected cancer (44,45,48–50,161), these separate studies have each focussed on a single symptom cohort. Ideally, research stratifying cancer risk should be repeated systematically for other non-specific symptoms, as this would enable their relative importance to be assessed, and provide the opportunity to explore a full range of possible two or three-way combinations of non-specific symptoms. This would require full cohort study designs, including landmark approaches(207); or time-varying-exposure approaches(208,209). In a landmark design, by incorporating multiple different snapshots of a patient's co-occurring symptoms throughout their health records, rather than those that co-occur at one specific 'snapshot' (i.e. around a symptom index date), this might address the sample size limitations that have so far restricted symptom combinations being examined in cohort studies such as mine. As a result, it might be possible to study risk associated

with rarer and more transient signs and symptoms, or multiple combinations of symptoms (e.g. co-occurring fatigue, weight loss and abdominal pain), although decisions around the length of the time window within which to search for co-occurring symptoms would still need to be made.

### *Risk prediction tools to support clinical practice*

My research aimed to describe risk in a symptomatic cohort and provide foundational, stratified risk information to support diagnostic guidelines (see Section 3.5.1), and is distinct from risk prediction tools that aim to help GPs assess cancer risk for the individual patients using a broad range of detailed information. Both types of studies can be seen as lying on a spectrum, where the addition of increasingly complex information to a risk stratification study such as mine moves it closer to becoming a risk prediction tool. Therefore, the recommendations in the previous section that risk stratification in patients with fatigue (or other non-specific cancer symptoms) could be improved by adding more information, such as blood test results, comorbidities, lifestyle factors etc, naturally leads to the question of whether it will be useful to GPs to generate risk prediction tools that specifically cater to patients presenting with vague symptoms.

In the case of fatigue, its overall predictive value for cancer in a population of primary care patients is dwarfed by the relative contributions of other alarm symptoms, when considered in prediction models, as indicated by the exclusion of fatigue from the QCancer risk prediction tool (44,45). Nevertheless, 96% of patients who present with fatigue do not have an alarm symptom for cancer recorded concurrently (Chapter 4). Given the frequency of fatigue in primary care, and the diagnostic challenge it presents in the absence of alarm symptoms, in practice, doctors still need to know the level of cancer risk, and the most likely cancer sites if cancer is present, to optimally support the diagnosis process for these patients. A prediction model specifically for patients with fatigue may not be practical, but models could be considered that are trained specifically on patients presenting with various vague symptoms in the absence of alarm symptoms for cancer.

Recently, modelling strategies have developed that can make fuller use of patients' rich EHR records, such as landmark models that incorporate past repeated measures of risk factors of interest into prediction of subsequent disease risk(207). Innovations in dynamic modelling designs (e.g. deep learning used by Placido et al(210)) also present data-driven ways to develop cancer risk prediction tools that make fuller use of patients' rich EHR records. Methods used by these studies can incorporate the timing of symptoms in patients' EHRs into vectorised space, hence they can incorporate patients' detailed disease, symptom, and test result trajectories. They could also incorporate flags that detect changes in individual patients' healthcare use, a concept suggested in some studies of pre-diagnostic healthcare use changes featured in the literature review in Chapter 2. As discussed, I did not include healthcare use measures in my studies, due to the methodological challenges involved in detecting statistically significant changes in healthcare use at an individual level, and because I deemed the generation of foundational evidence of disease risk in this cohort (stratified by age, sex, and other symptoms) to be higher priority. However, developing prediction tools that incorporate generalised measures of illness, such as detecting increases in healthcare use, could be particularly useful in identifying patients at increased risk of cancer or other serious illnesses, because many other diseases (e.g. Parkinson's disease, dementia) are characterised by the onset of non-specific symptoms in their early stages(211).

The proliferation of risk prediction tools for cancer (and other diseases) also presents challenges, in terms of their practical integration into GP software systems(212) and how they should be used alongside diagnostic referral guidelines (such as those published by the National Institute for Health and Care Excellence (NICE)), and cohort-based studies that stratify cancer risk in cohorts of patients with symptoms (or other clinical features), such as mine. More broadly, questions remain as to the role of NICE referral guidelines; currently, recommendations are stratified by selected characteristics, such as age, sex, and other presenting symptoms or blood test results(16). These overall simple recommendations could become increasingly complex as more cohort studies and risk prediction tools are published that stratify cancer risk at increasingly granular scales (e.g. multiple combinations of symptoms and test results).

## 7.6 Conclusions

For the first time, my studies quantify short term cancer risk in patients presenting to their GP with new-onset fatigue, according to patient age and sex, and co-occurring symptoms, and contextualise such risk relative to the risk of other possible diagnoses. Strong arguments can be made for recommendations to prioritise investigating cancer in older men presenting with fatigue, for whom cancer is a relatively common diagnosis, though in both men and women, the presence of other certain vague symptoms elevates cancer risk above referral thresholds in the UK. The presence of fatigue alone does not provide conclusive information about the likely cancer site.

Future research could further support diagnostic guidelines by stratifying risk of cancer (and other diseases) in patients with fatigue (and other non-specific symptoms), according to other features such as the results of diagnostic investigations or presence of comorbidities. In particular, the recent development of multicancer early detection (MCED) tests holds promise for early cancer detection in patients presenting with non-specific symptoms such as fatigue. Research will be needed into the performance of these tests in such cohorts, and the added information that positive and negative test results can provide about patients' absolute risk of specific cancer sites. This could help to reclassify patients into higher (with a positive MCED test result) or lower (with a negative result) risk compared to current referral thresholds.

My studies demonstrate the great potential of evidence from electronic health records (EHRs) to inform the continued development of referral guidelines for suspected cancer, as well as broader diagnostic guidelines for a range of diseases in symptomatic patients. EHRs offer unparalleled scale and population coverage in the UK, although they continue to present methodological challenges; namely the selective recording of symptoms in coded data and loss to follow up in UK primary care EHR databases, and 'messy' data items that need careful phenotyping.

Recent efforts provide exciting opportunities to maximise the richness of EHRs to generate better information about patients' risk of cancer and other diseases. Improvements in developing and sharing phenotypes now enable studies to contextualise cancer risk alongside a large number of possible disease outcomes, of which my study is one of the first, and the only to have used a data-driven approach to assess hundreds of diseases.

While my research aimed to offer foundational cancer risk stratification in an understudied group of patients presenting with fatigue, there are also developments in the area risk prediction tools that could support GPs to detect cancer (and other diseases) in patients presenting with non-specific symptoms, by leveraging the full detail of patients' longitudinal records (such as landmarking and dynamic modelling). These could include the development of risk prediction tools that incorporate general changes in patients' health and morbidity trajectories (e.g. changing healthcare use, blood test results, accrual of diagnoses, etc.). Questions remain as to how such tools can be developed, and subsequently embedded into policy and implemented in practice, alongside existing or as part of future diagnostic guidelines. Together, these efforts could help detect cancer earlier in primary care, for which there is great potential, as shown by population level changes in healthcare use long before cancer diagnosis. By detecting cancers earlier, patients' survival and quality of life could be much improved.

# 8. Chapter 8: Personal development and contributions

## 8.1 Publications

*Articles published during the Thesis study period (since September 2019)*

### 8.1.1 Thesis publications

White, B., Renzi, C., Rafiq, M., Zakkak, N., Gonzalez-Izquierdo, A., Denaxas, S., Nicholson, B., Lyratzopoulos, G., Barclay, M. Incident disease among patients presenting in primary care with fatigue: a population-based cohort study. In preparation.

White, B., Renzi, C., Barclay, M., & Lyratzopoulos, G. (2023). Underlying cancer risk among patients with fatigue and other vague symptoms: a population-based cohort study in primary care. British Journal of General Practice, BJGP.2022.0371. https://doi.org/10.3399/BJGP.2022.0371

White, B., Rafiq, M., Gonzalez-Izquierdo, A., Hamilton, W., Price, S., & Lyratzopoulos, G. (2022). Risk of cancer following primary care presentation with fatigue: a population-based cohort study of a quarter of a million patients. British Journal of Cancer, 126(11), 1627–1636. https://doi.org/10.1038/s41416-022-01733-6

White, B., Renzi, C., Rafiq, M., Abel, G. A., Jensen, H., & Lyratzopoulos, G. (2022). Does changing healthcare use signal opportunities for earlier detection of cancer? A review of studies using information from electronic patient records. Cancer Epidemiology, 76, 102072. https://doi.org/10.1016/j.canep.2021.102072

### 8.1.2 Related publications

*Relating to my job role, supervision of other PhD students, or previous research projects*

Benitez Majano, S., Lyratzopoulos, G., de Wit, N. J., White, B., Rachet, B., Helsper, C., Usher-Smith, J., & Renzi, C. (2022). Mental Health Morbidities and Time to Cancer Diagnosis Among Adults With Colon Cancer in England. JAMA Network Open, 5(10), e2238569. https://doi.org/10.1001/JAMANETWORKOPEN.2022.38569

Whitfield, E., White, B., Denaxas, S., & Lyratzopoulos, G. (2023). Diagnostic windows in non-neoplastic diseases: A systematic review. British Journal of General Practice, BJGP.2023.0044. https://doi.org/10.3399/BJGP.2023.0044

Fry, A., White, B., Nagarwalla, D., Shelton, J., & Jack, R. H. (2023). Relationship between ethnicity and stage at diagnosis in England: a national analysis of six cancer sites. BMJ Open, 13(1), e062079. https://doi.org/10.1136/BMJOPEN-2022-062079

White, B., Nordin, A., Fry, A., Ahmad, A., McPhail, S., Roe, C., Rous, B., Smittenaar, R., & Shelton, J. (2019). Geographic variation in the use of lymphadenectomy and external-beam radiotherapy for endometrial cancer: a cross-sectional analysis of population-based data. BJOG: An International

Journal of Obstetrics & Gynaecology, 126(12), 1456–1465. https://doi.org/10.1111/1471-0528.15914

## 8.2 Contributions to wider research community

To support future research and study replication, data management and analysis code, and Read code lists used to conduct the empirical studies in Chapter 5 and 6 are available online at https://github.com/rmjlrwh/Fatigue and https://github.com/rmjlrwh/FatigueRiskMap

I recorded a podcast with BJGP to accompany the publication of the empirical study in Chapter 5 (24th January 2023). "Episode 102: Combining vague cancer symptoms to improve referrals for suspected cancer". Available at https://bjgplife.com/episode-102-combining-vague-cancer-symptoms-to-improve-referrals-for-suspected-cancer/

I recorded an interview for Imperial University's science comms radio programme (""), alongside my colleague Emma Whitfield, to accompany the publication of the Whitfield et al (2023) publication. In preparation.

I delivered training talks to Cancer Research UK and the CanTest collaboration, as well as presenting my work to collaborators in the International Alliance for Cancer Early Detection (ACED), the THINK database research group at UCL, and the CanTest 'Summer school' and related events.

I reviewed 4 submitted manuscripts, including 2 for BJGP journal, 1 for Cancer Epidemiology, and 1 for Scandinavian Journal of Primary Healthcare.

## 8.3 Conferences attended

My record of conference attendance was affected by travel restrictions and conference cancellations introduced during the Covid pandemic, which commenced 6 months after my PhD enrolment. I was, however, able to attend the following conferences:

American Medical Informatics Association (AMIA), November 2022, Washington DC

World Cancer Conference (WCC), October 2022, Geneva

Health Data Research (HDR) UK Scientific Conference, December 2022, virtual

Cancer Research UK (CRUK) Early Detection of Cancer Conference, October 2020, virtual

## 8.4 Training attended

My attendance of training courses was affected by remote working patterns and the restricted offer of courses due to the Covid pandemic, which commenced 6 months after my PhD enrolment. I was, however, able to attend the following:

Royal Statistical Society, 3-4th October 2022. Survival analysis.

Keele university, 14-16th September 2021. Statistical methods for risk prediction & prognostic models.

ACED (International Alliance for Cancer Early Detection) -MCRC (Manchester Cancer Research Centre) Workshop, 13th October 2020. The use of routine healthcare data for research on cancer early detection.

Great Ormond Street Institute of Child Health, 8-9th January 2020. Introduction to Hospital Episode Statistics.

UCL THINK, 19-20th November 2019. Introduction to Primary Care Databases.

National Centre for Research Methods (NCRM), 5-6th November 2019, SQL for Biomedical Researchers.

## 8.5 'On the job' experience and training

I developed skills in scientific writing, including submitting scientific publications and addressing reviewer feedback, as well as reviewing scientific publications. This related to both literature review skills, and a range of competencies and knowledge involved in the conduct of electronic health records research.

As part of my PhD, I learnt to manage data, conduct statistical analysis, and visualise data using R. I improved my existing proficiency with SQL and Stata. Through my management of the datasets, and through professional socialisation with senior data managers at UCL's Institute of Health Informatics (IHI), I also developed an understanding of how to set up a SQL server, including computing resource management, setting up ODBC services, and using Putty to manage data files.

I was also closely involved in developing research protocols and data specifications when applying for access to data from CPRD, and managed the data life cycle from applying for, receiving, storing, publishing, and deleting EHR data. As part of this, I supervised another PhD student (Emma Whitfield) in developing her data protocol, and storage/ management of her data.

# 9. References


1. Pottegård A, Hallas J. New use of prescription drugs prior to a cancer diagnosis. Pharmacoepidemiol Drug Saf. 2017 Feb 1;26(2):223–7. Available from: http://doi.wiley.com/10.1002/pds.4145

2. Ewing M, Naredi P, Zhang C, Lindsköld L, Månsson J. Clinical features of patients with non-metastatic lung cancer in primary care: a case-control study. BJGP Open. 2018 Apr 1;2(1):bjgpopen18X101397. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30564706

3. Tørring ML, Falborg AZ, Jensen H, Neal RD, Weller D, Reguilon I, et al. Advanced-stage cancer and time to diagnosis: An International Cancer Benchmarking Partnership (ICBP) cross-sectional study. Eur J Cancer Care (Engl). 2019 Sep 22;28(5). Available from: https://pubmed.ncbi.nlm.nih.gov/31119836/

4. Neal RD, Tharmanathan P, France B, Din NU, Cotton S, Fallon-Ferguson J, et al. Is increased time to diagnosis and treatment in symptomatic cancer associated with poorer outcomes? Systematic review [Internet]. Vol. 112, British Journal of Cancer. Springer Nature; 2015. p. S92–107. Available from: https://pubmed.ncbi.nlm.nih.gov/25734382/

5. Mendonca SC, Abel GA, Saunders CL, Wardle J, Lyratzopoulos G. Pre-referral general practitioner consultations and subsequent experience of cancer care: evidence from the English Cancer Patient Experience Survey. Eur J Cancer Care (Engl). 2016 May 1;25(3):478–90. Available from: http://doi.wiley.com/10.1111/ecc.12353

6. Richards MA. The National Awareness and Early Diagnosis Initiative in England: Assembling the evidence. Br J Cancer. 2009 Dec 3;101(Suppl 2):S1–4.

7. Lyratzopoulos G, Abel GA. Assessing patients at risk of symptomatic-but-as-yet-undiagnosed cancer in primary care using information from patient records. Br J Cancer. 2020 Apr 15;1–3.

8. Hansen PL, Hjertholm P, Vedsted P. Increased diagnostic activity in general practice during the year preceding colorectal cancer diagnosis. Int J Cancer. 2015 Aug 1;137(3):615–24. Available from: http://doi.wiley.com/10.1002/ijc.29418

9. Nygaard C, Jensen H, Christensen J, Vedsted P. Health care use before a diagnosis of primary intracranial tumor: a Danish nationwide register study. Clin Epidemiol. 2018 Jul;Volume 10:809–29. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30038522

10. Koshiaris C, Van den Bruel A, Oke JL, Nicholson BD, Shephard E, Braddick M, et al. Early detection of multiple myeloma in primary care using blood tests: a case–control study in primary care. British Journal of General Practice. 2018 Sep 1;68(674):e586–93. Available from: http://bjgp.org/lookup/doi/10.3399/bjgp18X698357

11. Lyratzopoulos G, Vedsted P, Singh H. Understanding missed opportunities for more timely diagnosis of cancer in symptomatic patients after presentation. Vol. 112, British journal of cancer. 2015. p. S84–91.

12. Zhou Y, Mendonca SC, Abel GA, Hamilton W, Walter FM, Johnson S, et al. Variation in 'fast-track' referrals for suspected cancer by patient characteristic and cancer diagnosis: evidence from 670 000 patients with cancers of 35 different sites. Br J Cancer. 2018 Jan 28;118(1):24–31. Available from: http://www.nature.com/articles/bjc2017381

13. Din NU, Ukoumunne OC, Rubin G, Hamilton W, Carter B, Stapley S, et al. Age and Gender Variations in Cancer Diagnostic Intervals in 15 Cancers: Analysis of Data from the UK Clinical Practice Research Datalink. Katoh M, editor. PLoS One. 2015 May 15;10(5):e0127717. Available from: https://dx.plos.org/10.1371/journal.pone.0127717

14. Lyratzopoulos G, Abel GA, McPhail S, Neal RD, Rubin GP. Measures of promptness of cancer diagnosis in primary care: Secondary analysis of national audit data on patients with 18 common and rarer cancers. Br J Cancer. 2013 Feb 19;108(3):686–90. Available from: /pmc/articles/PMC3593564/?report=abstract

15. Lyratzopoulos G, Saunders CL, Abel GA, McPhail S, Neal RD, Wardle J, et al. The relative length of the patient and the primary care interval in patients with 28 common and rarer cancers. Br J Cancer. 2015 Mar 3;112 Suppl(Suppl 1).

16. National Institute for Health and Care Excellence. Suspected cancer: recognition and referral NICE Guideline [NG12] [Internet]. NICE; 2015. Available from: https://www.nice.org.uk/guidance/ng12/chapter/Recommendations-organised-by-symptom-and-findings-of-primary-care-investigations

17. Jensen H, Tørring ML, Olesen F, Overgaard J, Vedsted P. Cancer suspicion in general practice, urgent referral and time to diagnosis: A population-based GP survey and registry study. BMC Cancer. 2014 Aug 30;14(1).

18. Pearson C, Poirier V, Fitzgerald K, Rubin G, Hamilton W. Cross-sectional study using primary care and cancer registration data to investigate patients with cancer presenting with non-specific symptoms. BMJ Open. 2020 Jan 1;10(1):e033008. Available from: https://bmjopen.bmj.com/content/10/1/e033008

19. Rubin G, Berendsen A, Crawford SM, Dommett R, Earle C, Emery J, et al. The expanding role of primary care in cancer control. Lancet Oncol. 2015 Sep 1;16(12):1231–72.

20. Green T, Atkin K, Macleod U. Cancer detection in primary care: insights from general practitioners. British Journal of Cancer 2015 112:1. 2015 Mar 3;112(1):S41–9. Available from: https://www.nature.com/articles/bjc201541

21. Nicholson K, Stewart M, Thind A. Examining the symptom of fatigue in primary care: A comparative study using electronic medical records. J Innov Health Inform. 2015 Jan 1;22(1):235–43. Available from: http://dx.doi.org/10.14236/jhi.v22i1.91

22. Hamilton W, Watson J, Round A. Rational testing: Investigating fatigue in primary care. BMJ (Online). 2010 Sep 4;341(7771):502–4. Available from: https://www.bmj.com/content/341/bmj.c4259

23. Cathébras PJ, Robbins JM, Kirmayer LJ, Hayton BC. Fatigue in primary care - Prevalence, psychiatric comorbidity, illness behavior, and outcome. J Gen Intern Med.

1992 May;7(3):276–86. Available from: https://link.springer.com/article/10.1007/BF02598083

24. Cullen W, Kearney Y, Bury G. Prevalence of fatigue in general practice. Ir J Med Sci. 2002;171(1):10–2. Available from: https://link.springer.com/article/10.1007/BF03168931

25. McAteer A, Elliott AM, Hannaford PC. Ascertaining the size of the symptom iceberg in a UK-wide community-based survey. British Journal of General Practice. 2011 Jan;61(582):e1. Available from: /pmc/articles/PMC3020067/

26. Hannaford PC, Thornton AJ, Murchie P, Whitaker KL, Adam R, Elliott AM. Patterns of symptoms possibly indicative of cancer and associated help-seeking behaviour in a large sample of United Kingdom residents—The USEFUL study. Bowles E, editor. PLoS One. 2020 Jan 24;15(1):e0228033. Available from: https://dx.plos.org/10.1371/journal.pone.0228033

27. Elnegaard S, Andersen RS, Pedersen AF, Larsen PV, Søndergaard J, Rasmussen S, et al. Self-reported symptoms and healthcare seeking in the general population -exploring "the Symptom Iceberg." BMC Public Health. 2015 Jul 21;15(1):1–11. Available from: https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-015-2034-5

28. National Institute for Health and Care Excellence. The recognition and initial management of ovarian cancer: evidence review [Internet]. 2011. Available from: https://www.nice.org.uk/guidance/cg122/evidence/evidence-review-pdf-181688798

29. National Institute for Health and Care Excellence. Suspected cancer: recognition and referral NICE Guideline [NG12]. Appendix F: Evidence Review [Internet]. 2015. Available from: https://www.nice.org.uk/guidance/ng12/evidence/full-guideline-pdf-2676000277

30. Koo MM, Swann R, McPhail S, Abel GA, Elliss-Brookes L, Rubin GP, et al. Presenting symptoms of cancer and stage at diagnosis: evidence from a cross-sectional, population-based study. Lancet Oncol. 2020 Jan 1;21(1):73–9.

31. Koo MM, Hamilton W, Walter FM, Rubin GP, Lyratzopoulos G. Symptom Signatures and Diagnostic Timeliness in Cancer Patients: A Review of Current Evidence [Internet]. Vol. 20, Neoplasia (United States). Neoplasia Press, Inc.; 2018. p. 165–74. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1476558617304268

32. Friedman GD, Skilling JS, Udaltsova N V., Smith LH. Early symptoms of ovarian cancer: A case-control study without recall bias. Fam Pract. 2005;22(5):548–53.

33. Dommett RM, Redaniel T, Stevens MC, Martin RM, Hamilton W. Risk of childhood cancer with symptoms in primary care: a population-based case-control study. British Journal of General Practice. 2013 Jan 1;63(606):e22–9. Available from: https://pubmed.ncbi.nlm.nih.gov/23336454/

34. Dommett RM, Redaniel MT, Stevens MCG, Hamilton W, Martin RM. Features of cancer in teenagers and young adults in primary care: a population-based nested case–control study. Br J Cancer. 2013 Jun 25;108(11):2329–33. Available from: https://www.nature.com/articles/bjc2013191

35. Attanucci CA, Ball HG, Zweizig SL, Chen AH. Differences in symptoms between patients with benign and malignant ovarian neoplasms. Am J Obstet Gynecol. 2004;190(5):1435–7.

36. Hamilton W, Peters TJ, Round A, Sharp D. What are the clinical features of lung cancer before the diagnosis is made? A population based case-control study. Thorax. 2005 Dec;60(12):1059–65.

37. Favrat B, Cornuz J. BMJ Best Practice. 2022. Assessment of fatigue. Available from: https://bestpractice.bmj.com/topics/en-gb/571

38. Hamilton W. The CAPER studies: five case-control studies aimed at identifying and quantifying the risk of cancer in symptomatic primary care patients. Br J Cancer. 2009 Dec 3;101(S2):S80–6. Available from: http://dx.doi.org/10.1038/sj.bjc.6605396

39. Cornuz J, Guessous I, Favrat B. Fatigue: A practical approach to diagnosis in primary care. CMAJ. 2006 Mar 14;174(6):765–7. Available from: https://www.cmaj.ca/content/174/6/765

40. Sharpe M, Wilks D. Fatigue. BMJ. 2002 Aug 31;325(7362):480. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1124000/

41. Simon C. Tiredness, Fatigue and Lethargy. InnovAiT: Education and inspiration for general practice. 2008 Mar 1;1(3):199–205. Available from: https://journals.sagepub.com/doi/10.1093/innovait/inn017

42. Stadje R, Dornieden K, Baum E, Becker A, Biroga T, Bösner S, et al. The differential diagnosis of tiredness: A systematic review. BMC Fam Pract. 2016 Oct 20;17(1):147. Available from: http://bmcfampract.biomedcentral.com/articles/10.1186/s12875-016-0545-5

43. National Institute for Health and Care Excellence. Recommendations organised by symptom and findings of primary care investigations, suspected cancer: recognition and referral [Internet]. NICE; 2015. Available from: https://www.nice.org.uk/guidance/ng12/chapter/Recommendations-organised-by-symptom-and-findings-of-primary-care-investigations

44. Hippisley-Cox J, Coupland C. Symptoms and risk factors to identify women with suspected cancer in primary care: Derivation and validation of an algorithm. British Journal of General Practice. 2013 Jan 1;63(606):e11–21. Available from: https://bjgp.org/content/63/606/e11

45. Hippisley-Cox J, Coupland C. Symptoms and risk factors to identify men with suspected cancer in primary care: Derivation and validation of an algorithm. British Journal of General Practice. 2013 Jan;63(606):e1. Available from: /pmc/articles/PMC3529287/?report=abstract

46. NHS England, NHS Improvement. Rapid Diagnostic Centres: Vision and 2019/20 Implementation Specification [Internet]. 2019. Available from: https://www.england.nhs.uk/wp-content/uploads/2019/07/rdc-vision-and-1920-implementation-specification.pdf

47. Erridge S, Lyratzopoulos G, Renzi C, Millar A, Lee R. Rapid Diagnostic Centres and early cancer diagnosis. British Journal of General Practice. 2021 Nov 1;71(712):487–8. Available from: https://bjgp.org/content/71/712/487

48. Price SJ, Gibson N, Hamilton WT, King A, Shephard EA. Intra-abdominal cancer risk with abdominal pain: a prospective cohort primary care study. British Journal of General Practice. 2022 May 16;72(718):e361–8. Available from: https://bjgp.org/content/early/2022/04/04/BJGP.2021.0552

49. Nicholson BD, Aveyard P, Price SJ, Hobbs FR, Koshiaris C, Hamilton W. Prioritising primary care patients with unexpected weight loss for cancer investigation: Diagnostic accuracy study. The BMJ. 2020 Aug 13;370. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7424394/

50. Herbert A, Rafiq M, Pham TM, Renzi C, Abel GA, Price S, et al. Predictive values for different cancers and inflammatory bowel disease of 6 common abdominal symptoms among more than 1.9 million primary care patients in the UK: A cohort study. Basu S, editor. PLoS Med. 2021 Aug 2;18(8):e1003708. Available from: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003708

51. Rasmussen S, Haastrup PF, Balasubramaniam K, Elnegaard S, Christensen R dePont, Storsveen MM, et al. Predictive values of colorectal cancer alarm symptoms in the general population: a nationwide cohort study. Br J Cancer. 2019 Mar 22;120(6):595–600. Available from: /pmc/articles/PMC6461905/

52. Holtedahl K, Hjertholm P, Borgquist L, Donker GA, Buntinx F, Weller D, et al. Abdominal symptoms and cancer in the abdomen: prospective cohort study in European primary care. British Journal of General Practice. 2018 May 1;68(670):e301–10. Available from: /pmc/articles/PMC5916077/

53. Wilson J, Morgan S, Parker M, van Driel M. Vol. 43, Australian Family Physician. The Royal Australian College of General Practitioners; 2014. p. 457–61 RACGP - Fatigue – a rational approach to investigation. Available from: https://www.racgp.org.au/afp/2014/july/fatigue/

54. National Institute for Health and Care Excellence. Clinical Knowledge Summary: Tiredness/ fatigue in adults: Diagnosis [Internet]. 2021. Available from: https://cks.nice.org.uk/topics/tiredness-fatigue-in-adults/diagnosis/assessment/

55. Clarke C, Hamilton W, Price S, Bailey SER. Association of non-malignant diseases with thrombocytosis: a prospective cohort study in general practice. British Journal of General Practice. 2020 Dec 1;70(701):e852–7. Available from: https://bjgp.org/content/70/701/e852

56. Tørring ML, Frydenberg M, Hamilton W, Hansen RP, Lautrup MD, Vedsted P. Diagnostic interval and mortality in colorectal cancer: U-shaped association demonstrated for three different datasets. J Clin Epidemiol. 2012 Jun;65(6):669–78. Available from: https://linkinghub.elsevier.com/retrieve/pii/S089543561100388X

57. Tørring ML, Murchie P, Hamilton W, Vedsted P, Esteva M, Lautrup M, et al. Evidence of advanced stage colorectal cancer with longer diagnostic intervals: A pooled analysis of seven primary care cohorts comprising 11 720 patients in five countries. Br J

Cancer. 2017 Sep 5;117(6):888–97. Available from: https://pubmed.ncbi.nlm.nih.gov/28787432/

58. Crosby D, Lyons N, Greenwood E, Harrison S, Hiom S, Moffat J, et al. A roadmap for the early detection and diagnosis of cancer [Internet]. Vol. 21, The Lancet Oncology. Lancet Publishing Group; 2020. p. 1397–9. Available from: https://pubmed.ncbi.nlm.nih.gov/33031732/

59. Cancer Research UK. Early Detection and Diagnosis of Cancer: A Roadmap to the Future [Internet]. 2020. Available from: https://www.cancerresearchuk.org/sites/default/files/early_detection_diagnosis_of_cancer_roadmap.pdf

60. Groome PA, Webber C, Whitehead M, Moineddin R, Grunfeld E, Eisen A, et al. Determining the Cancer Diagnostic Interval Using Administrative Health Care Data in a Breast Cancer Cohort. JCO Clin Cancer Inform. 2019 Dec;3(3):1–10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/31112418

61. Kuiper JG, Van Herk-Sukel MPP, Lemmens VEPP, Kuipers EJ, Herings RMC. A steep increase in healthcare seeking behaviour in the last months before colorectal cancer diagnosis. BMC Fam Pract. 2021;22:121. Available from: https://doi.org/10.1186/s12875-021-01482-0

62. Jessen NH, Jensen H, Falborg AZ, Glerup H, Gronbaek H, Vedsted P. Abdominal investigations in the year preceding a diagnosis of abdominal cancer: A register-based cohort study in Denmark. Cancer Epidemiol. 2021 Jun 1;72:101926. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1877782121000436

63. Guldbrandt LM, Møller H, Jakobsen E, Vedsted P. General practice consultations, diagnostic investigations, and prescriptions in the year preceding a lung cancer diagnosis. Cancer Med. 2017 Jan 1;6(1):79–88. Available from: http://doi.wiley.com/10.1002/cam4.965

64. Hauswaldt J, Hummers-Pradier E, Himmel W. Does an increase in visits to general practice indicate a malignancy? BMC Fam Pract. 2016 Jul 26;17(1):94. Available from: http://bmcfampract.biomedcentral.com/articles/10.1186/s12875-016-0477-0

65. Ahrensberg JM, Fenger-Grøn M, Vedsted P. Use of Primary Care during the Year before Childhood Cancer Diagnosis: A Nationwide Population-Based Matched Comparative Study. Miller TW, editor. PLoS One. 2013 Mar 12;8(3):e59098. Available from: https://dx.plos.org/10.1371/journal.pone.0059098

66. Ewing M, Naredi P, Nemes S, Zhang C, Månsson J. Increased consultation frequency in primary care, a risk marker for cancer: a case–control study. Scand J Prim Health Care. 2016 Apr 2;34(2):204–11. Available from: https://pubmed.ncbi.nlm.nih.gov/27189513/

67. Zhou Y, Abel GA, Hamilton W, Singh H, Walter FM, Lyratzopoulos G. Imaging activity possibly signalling missed diagnostic opportunities in bladder and kidney cancer: A longitudinal data-linkage study using primary care electronic health records. Cancer Epidemiol. 2020 Jun 1;66:101703. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1877782120300370

68. Rædkjær M, Maretty-Kongstad K, Baad-Hansen T, Safwat A, Petersen MM, Keller J, et al. Use of Healthcare Services Two Years before Diagnosis in Danish Sarcoma Patients, 2000–2013. Sarcoma. 2019 May 7;2019:1–10. Available from: https://www.hindawi.com/journals/sarcoma/2019/8108590/

69. Renzi C, Lyratzopoulos G, Hamilton W, Rachet B. Opportunities for reducing emergency diagnoses of colon cancer in women and men: A data-linkage study on pre-diagnostic symptomatic presentations and benign diagnoses. Eur J Cancer Care (Engl). 2019 Mar 1;28(2):1–13.

70. Renzi C, Lyratzopoulos G, Hamilton W, Maringe C, Rachet B. Contrasting effects of comorbidities on emergency colon cancer diagnosis: a longitudinal data-linkage study in England. BMC Health Serv Res. 2019 Dec 15;19(1):311. Available from: https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-019-4075-4

71. Jensen H, Vedsted P, Møller H. Consultation frequency in general practice before cancer diagnosis in relation to the patient's usual consultation pattern: A population-based study. Cancer Epidemiol. 2018 Aug 1;55:142–8. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1877782118302819

72. Chu TPCC, Shah A, Walker D, Coleman MP. How Do Biological Characteristics of Primary Intracranial Tumors Affect Their Clinical Presentation in Children and Young Adults? J Child Neurol. 2018 Jul 3;33(8):503–11. Available from: http://journals.sagepub.com/doi/10.1177/0883073818767562

73. Chu TPC, Shah A, Walker D, Coleman MP. Where are the opportunities for an earlier diagnosis of primary intracranial tumours in children and young adults? European Journal of Paediatric Neurology. 2017 Mar 1;21(2):388–95. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1090379816301945

74. Mcdonald L, Carroll R, Harish A, Tanna N, Mehmud F, Alikhan R, et al. Suspected cancer symptoms and blood test results in primary care before a diagnosis of lung cancer: a case–control study. Future Oncology. 2019 Nov 31;15(33):3755–62. Available from: https://www.futuremedicine.com/doi/10.2217/fon-2019-0442

75. Christensen KG, Fenger-Grøn M, Flarup KR, Vedsted P. Use of general practice, diagnostic investigations and hospital services before and after cancer diagnosis - a population-based nationwide registry study of 127,000 incident adult cancer patients. BMC Health Serv Res. 2012;12:224. Available from: https://pubmed.ncbi.nlm.nih.gov/22838741/

76. Chu TPCC, Shah A, Walker D, Coleman MP. Pattern of symptoms and signs of primary intracranial tumours in children and young adults: a record linkage study. Arch Dis Child. 2015 Dec 1;100(12):1115–22. Available from: http://adc.bmj.com/lookup/doi/10.1136/archdischild-2014-307578

77. Morrell S, Young J, Roder D. The burden of cancer on primary and secondary health care services before and after cancer diagnosis in New South Wales, Australia. BMC Health Serv Res. 2019 Jun 27;19(1). Available from: https://pubmed.ncbi.nlm.nih.gov/31248405/

78. Wang Y, Freemantle N, Nazareth I, Hunt K. Gender Differences in Survival and the Use of Primary Care Prior to Diagnosis of Three Cancers: An Analysis of Routinely Collected UK General Practice Data. Katoh M, editor. PLoS One. 2014 Jul 11;9(7):e101562. Available from: https://dx.plos.org/10.1371/journal.pone.0101562

79. Renzi C, Lyratzopoulos G, Card T, Chu TPC, Macleod U, Rachet B. Do colorectal cancer patients diagnosed as an emergency differ from non-emergency patients in their consultation patterns and symptoms? A longitudinal data-linkage study in England. Br J Cancer. 2016 Sep 27;115(7):866–75. Available from: http://dx.doi.org/10.1038/bjc.2016.250

80. Friis Abrahamsen C, Ahrensberg JM, Vedsted P. Utilisation of primary care before a childhood cancer diagnosis: do socioeconomic factors matter?: A Danish nationwide population-based matched cohort study. BMJ Open. 2018 Aug 17;8(8):e023569. Available from: https://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2018-023569

81. Ades AE, Biswas M, Welton NJ, Hamilton W. Symptom lead time distribution in lung cancer: Natural history and prospects for early diagnosis. Int J Epidemiol. 2014 Dec 1;43(6):1865–73.

82. Ansell P, Johnston T, Simpson J, Crouch S, Roman E, Picton S. Brain Tumor Signs and Symptoms: Analysis of Primary Health Care Records From the UKCCS. Pediatrics. 2010 Jan 1;125(1):112–9. Available from: http://pediatrics.aappublications.org/cgi/doi/10.1542/peds.2009-0254

83. Ahrensberg JM, Fenger-Grøn M, Vedsted P. Primary Care Use before Cancer Diagnosis in Adolescents and Young Adults – A Nationwide Register Study. Tsokos M, editor. PLoS One. 2016 May 20;11(5):e0155933. Available from: https://dx.plos.org/10.1371/journal.pone.0155933

84. Morris M, Woods LM, Bhaskaran K, Rachet B. Do pre-diagnosis primary care consultation patterns explain deprivation-specific differences in net survival among women with breast cancer? An examination of individually-linked data from the UK West Midlands cancer registry, national screening programme. BMC Cancer. 2017 Feb 23;17(1).

85. Schuemie MJ, Ryan PB, Man KKC, Wong ICK, Suchard MA, Hripcsak G. A plea to stop using the case-control design in retrospective database studies. Stat Med. 2019 Sep 30;38(22):4199–208. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8215

86. Weller D, Vedsted P, Rubin G, Walter FM, Emery J, Scott S, et al. The Aarhus statement: improving design and reporting of studies on early cancer diagnosis. Br J Cancer. 2012 Mar 13;106(7):1262–7. Available from: http://www.nature.com/articles/bjc201268

87. Rutjes AWS, Reitsma JB, Di Nisio M, Smidt N, Van Rijn JC, Bossuyt PMM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ : Canadian Medical Association Journal. 2006 Feb 14;174(4):469. Available from: /pmc/articles/PMC1373751/

88. Kostopoulou O, Tracey C, Delaney BC. Can decision support combat incompleteness and bias in routine primary care data? Journal of the American Medical Informatics Association. 2021 Mar 11; Available from: https://academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocab025/6168487

89. Price SJ, Stapley SA, Shephard E, Barraclough K, Hamilton WT. Is omission of free text records a possible source of data loss and bias in Clinical Practice Research Datalink studies? A case-control study. BMJ Open. 2016 May 1;6(5):e011664.

90. Girolamo C Di, Walters S, Gildea C, Majano SB, Rachet B, Morris M. Can we assess Cancer Waiting Time targets with cancer survival? A population-based study of individually linked data from the National Cancer Waiting Times monitoring dataset in England, 2009-2013 [Internet]. Vol. 13, PLoS ONE. Public Library of Science; 2018. Available from: /pmc/articles/PMC6104918/?report=abstract

91. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. Journal of the American Medical Informatics Association. 2017 Jan 1;24(1):198–208. Available from: /pmc/articles/PMC5201180/

92. Ingebrigtsen SG, Scheel BI, Hart B, Thorsen T, Holtedahl K. Frequency of 'warning signs of cancer' in Norwegian general practice, with prospective recording of subsequent cancer. Fam Pract. 2013 Apr 1;30(2):153–60. Available from: https://academic.oup.com/fampra/article/30/2/153/499494

93. Lawrenson R, Logie J, Marks C. Risk of colorectal cancer in general practice patients presenting with rectal bleeding, change in bowel habit or anaemia. Eur J Cancer Care (Engl). 2006 Jul;15(3):267–71. Available from: https://pubmed.ncbi.nlm.nih.gov/16882123/

94. Watson J, Nicholson BD, Hamilton W, Price S. Identifying clinical features in primary care electronic health record studies: methods for codelist development. BMJ Open. 2017 Nov 22;7(11):e019637. Available from: https://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2017-019637

95. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association. 2013 Jan 1;20(1):117–21. Available from: https://academic.oup.com/jamia/article/20/1/117/2909152

96. Withrow DR, Oke J, Smith CF, Hobbs R, Nicholson BD. Serious disease risk among patients with unexpected weight loss: a matched cohort of over 70 000 primary care presentations. J Cachexia Sarcopenia Muscle. 2022 Sep 3; Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/jcsm.13056

97. Nicholson BD, Hamilton W, Koshiaris C, Oke JL, Hobbs FDR, Aveyard P. The association between unexpected weight loss and cancer diagnosis in primary care: a matched cohort analysis of 65,000 presentations. Br J Cancer. 2020 Jun 9;122(12):1848–56. Available from: http://www.nature.com/articles/s41416-020-0829-3

98. Price SJ, Gibson N, Hamilton WT, Bostock J, Shephard EA. Diagnoses after newly recorded abdominal pain in primary care: observational cohort study. British Journal

of General Practice. 2022 Aug 1;72(721):e564–70. Available from: /pmc/articles/PMC9242675/

99. Nicholson BD, Oke JL, Aveyard P, Hamilton WT, Hobbs FDR. Individual inflammatory marker abnormalities or inflammatory marker scores to identify primary care patients with unexpected weight loss for cancer investigation? Br J Cancer. 2021 Feb 9;1–3. Available from: http://www.nature.com/articles/s41416-021-01282-4

100. Watson J, Whiting P, Salisbury C, Banks J, Hamilton W. Raised inflammatory markers as a predictor of one-year mortality: a cohort study in primary care in the UK using electronic health record data. BMJ Open. 2020 Oct 15;10(10). Available from: https://pubmed.ncbi.nlm.nih.gov/33060080/

101. Watson J, Salisbury C, Banks J, Whiting P, Hamilton W. Predictive value of inflammatory markers for cancer diagnosis in primary care: a prospective cohort study using electronic health records. Br J Cancer. 2019;120(11):1045–51. Available from: http://dx.doi.org/10.1038/s41416-019-0458-x

102. Watson J, Salisbury C, Whiting P, Banks J, Pyne Y, Hamilton W. Added value and cascade effects of inflammatory marker tests in UK primary care: a cohort study from the Clinical Practice Research Datalink. British Journal of General Practice. 2019 Jul;69(684):e470–8. Available from: http://bjgp.org/lookup/doi/10.3399/bjgp19X704321

103. Ekstrand C, Bahmanyar S, Cherif H, Kieler H, Linder M. Cancer risk in patients with primary immune thrombocytopenia – A Swedish nationwide register study. Cancer Epidemiol. 2020 Dec 1;69:101806. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1877782120301405

104. Knight R, Walker V, Ip S, Cooper JA, Bolton T, Keene S, et al. Association of COVID-19 With Major Arterial and Venous Thrombotic Diseases: A Population-Wide Cohort Study of 48 Million Adults in England and Wales. Circulation. 2022 Sep 9;146(12):892. Available from: /pmc/articles/PMC9484653/

105. Ankus E, Price SJ, Ukoumunne OC, Hamilton W, Bailey SER. Cancer incidence in patients with a high normal platelet count: a cohort study using primary care data. Fam Pract. 2018 Dec 12;35(6):671–5. Available from: https://pubmed.ncbi.nlm.nih.gov/29659802/

106. LT M, W H, SE B. Cancer incidence following a high-normal platelet count: cohort study using electronic healthcare records from English primary care. Br J Gen Pract. 2020 Sep 1;70(698):E622–8. Available from: https://pubmed.ncbi.nlm.nih.gov/32719013/

107. Jones R, Latinovic R, Charlton J, Gulliford MC. Alarm symptoms in early diagnosis of cancer in primary care: Cohort study using General Practice Research Database. Br Med J. 2007 May 19;334(7602):1040–4.

108. Boennelykke A, Jensen H, Østgård LSG, Falborg AZ, Hansen AT, Christensen KS, et al. Cancer risk in persons with new-onset anaemia: a population-based cohort study in Denmark. BMC Cancer. 2022 Dec 1;22(1):1–13. Available from: https://bmccancer.biomedcentral.com/articles/10.1186/s12885-022-09912-7

109. SE B, OC U, EA S, W H. Clinical relevance of thrombocytosis in primary care: a prospective cohort study of cancer incidence using English electronic medical records and cancer registry data. Br J Gen Pract. 2017 Jun 1;67(659):e405–13. Available from: https://pubmed.ncbi.nlm.nih.gov/28533199/

110. Hopkins R, Bailey SE, Hamilton WT, Shephard EA. Microcytosis as a risk marker of cancer in primary care: a cohort study using electronic patient records. British Journal of General Practice. 2020 Jul 1;70(696):e457–62. Available from: https://pubmed.ncbi.nlm.nih.gov/32366530/

111. Huang SH, LePendu P, Iyer S V., Tai-Seale M, Carrell D, Shah NH. Toward personalizing treatment for depression: predicting diagnosis and severity. Journal of the American Medical Informatics Association. 2014 Nov 1;21(6):1069–75. Available from: https://academic.oup.com/jamia/article/21/6/1069/2909304

112. Riley RD, Hayden JA, Steyerberg EW, Moons KGM, Abrams K, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. PLoS Med. 2013 Feb 5;10(2):e1001380. Available from: http://www.progress-partnership.

113. Weiskopf NG, Dorr DA, Jackson C, Lehmann HP, Thompson CA. Healthcare utilization is a collider: an introduction to collider bias in EHR data reuse. Journal of the American Medical Informatics Association. 2023 Feb 8; Available from: https://academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocad013/7031302

114. Elliott AM, McAteer A, Hannaford PC. Revisiting the symptom iceberg in today's primary care: Results from a UK population survey. BMC Fam Pract. 2011 Dec 7;12(1):16. Available from: https://bmcfampract.biomedcentral.com/articles/10.1186/1471-2296-12-16

115. Austin PC, Lee DS, Fine JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. Circulation. 2016 Feb 9;133(6):601–9. Available from: https://www.ahajournals.org/doi/abs/10.1161/circulationaha.115.017719

116. Schmidt JCF, Lambert PC, Gillies CL, Sweeting MJ. Patterns of rates of mortality in the Clinical Practice Research Datalink. PLoS One. 2022 Aug 1;17(8):e0265709. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0265709

117. Crowther MJ, Lambert PC. Parametric multistate survival models: Flexible modelling allowing transition-specific distributions with application to estimating clinically useful measures of effect differences. Stat Med. 2017 Dec 20;36(29):4719–42. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/sim.7448

118. Lambert PC, Dickman PW, Rutherford MJ. Comparison of different approaches to estimating age standardized net survival. BMC Med Res Methodol. 2015 Aug 15;15(1):1–13. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-015-0057-3

119. Martin RM, Donovan JL, Turner EL, Metcalfe C, Young GJ, Walsh EI, et al. Effect of a Low-Intensity PSA-Based Screening Intervention on Prostate Cancer Mortality: The

CAP Randomized Clinical Trial. JAMA. 2018 Mar 6;319(9):883–95. Available from: https://jamanetwork.com/journals/jama/fullarticle/2673968

120. Menon U, Gentry-Maharaj A, Burnell M, Singh N, Ryan A, Karpinskyj C, et al. Ovarian cancer population screening and mortality after long-term follow-up in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. The Lancet. 2021 Jun 5;397(10290):2182–93. Available from: http://www.thelancet.com/article/S0140673621007315/fulltext

121. de Koning HJ, van der Aalst CM, de Jong PA, Scholten ET, Nackaerts K, Heuvelmans MA, et al. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. New England Journal of Medicine. 2020 Feb 6;382(6):503–13. Available from: https://www.nejm.org/doi/full/10.1056/nejmoa1911793

122. University of Leicester. InterPreT Cancer Survival web tool [Internet]. Available from: https://interpret.le.ac.uk/methods.php?ver=1.00

123. Arnold M, Rutherford MJ, Ferlay J, Phd S, Bray Phd F, Biostatistics ) ;, et al. Progress in cancer survival, mortality, and incidence in seven high-income countries 1995â€“2014 (ICBP SURVMARK-2): a population-based study. Lancet Oncology. 2019;20:1493–505. Available from: https://www.

124. Putter H, Fiocco M, Gekus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med. 2007 May 20;26(11):2389–430. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/sim.2712

125. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. J Am Stat Assoc. 1999 Jun;94(446):496. Available from: https://www.jstor.org/stable/2670170?origin=crossref

126. Lewis JD, Bilker WB, Weinstein RB, Strom BL. The relationship between time since registration and measured incidence rates in the General Practice Research Database. Pharmacoepidemiol Drug Saf. 2005 Jul;14(7):443–51.

127. Nicholson BD, Aveyard P, Hamilton W, Bankhead CR, Koshiaris C, Stevens S, et al. The internal validation of weight and weight change coding using weight measurement data within the UK primary care Electronic Health Record. Clin Epidemiol. 2019;11:145. Available from: /pmc/articles/PMC6354686/

128. Fillon M. Breast cancer recurrence risk can remain for 10 to 32 years. CA Cancer J Clin. 2022 May 1;72(3):197–9. Available from: https://onlinelibrary.wiley.com/doi/full/10.3322/caac.21724

129. Rassen JA, Bartels DB, Schneeweiss S, Patrick AR, Murk W. Measuring prevalence and incidence of chronic conditions in claims and electronic health record databases. Clin Epidemiol. 2018 Dec 17;11:1–15. Available from: https://www.dovepress.com/measuring-prevalence-and-incidence-of-chronic-conditions-in-claims-and-peer-reviewed-fulltext-article-CLEP

130. Watson J, Jones HE, Banks J, Whiting P, Salisbury C, Hamilton W. Use of multiple inflammatory marker tests in primary care: using Clinical Practice Research Datalink to evaluate accuracy. British Journal of General Practice. 2019 Jul;69(684):e462–9. Available from: https://bjgp.org/lookup/doi/10.3399/bjgp19X704309

131. Suissa S. Immortal Time Bias in Pharmacoepidemiology. Am J Epidemiol. 2008 Feb 15;167(4):492–9. Available from: https://academic.oup.com/aje/article/167/4/492/233064

132. Gallagher AM, Dedman D, Padmanabhan S, Leufkens HGM, de Vries F. The accuracy of date of death recording in the Clinical Practice Research Datalink GOLD database in England compared with the Office for National Statistics death registrations. Pharmacoepidemiol Drug Saf. 2019 May 1;28(5):563–9. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/pds.4747

133. Arhi CS, Bottle A, Burns EM, Clarke JM, Aylin P, Ziprin P, et al. Comparison of cancer diagnosis recording between the Clinical Practice Research Datalink, Cancer Registry and Hospital Episodes Statistics. Cancer Epidemiol. 2018;57(October):148–57. Available from: https://doi.org/10.1016/j.canep.2018.08.009

134. Margulis A V., Fortuny J, Kaye JA, Calingaert B, Reynolds M, Plana E, et al. Validation of Cancer Cases Using Primary Care, Cancer Registry, and Hospitalization Data in the United Kingdom. Epidemiology. 2018 Mar 1;29(2):308–13.

135. Lesko CR, Edwards JK, Cole SR, Moore RD, Lau B. When to Censor? Am J Epidemiol. 2018 Mar 1;187(3):623. Available from: /pmc/articles/PMC6248498/

136. Henson KE, Elliss-Brookes L, Coupland VH, Payne E, Vernon S, Rous B, et al. Data Resource Profile: National Cancer Registration Dataset in England. Int J Epidemiol. 2020 Feb 1;49(1):16–16h. Available from: https://academic.oup.com/ije/article/49/1/16/5476570

137. French J, Chen C, Henson K, Shand B, Ferris P, Pencheon J, et al. Identification of patient prescribing predicting cancer diagnosis using boosted decision trees. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag; 2019. p. 328–33.

138. Hripcsak G, Shang N, Peissig PL, Rasmussen L V., Liu C, Benoit B, et al. Facilitating phenotype transfer using a common data model. J Biomed Inform. 2019 Aug 1;96:103253.

139. Ostropolets A, Reich C, Ryan P, Weng C, Molinaro A, DeFalco F, et al. Characterizing database granularity using SNOMED-CT hierarchy. AMIA Annual Symposium Proceedings. 2020;2020:983. Available from: /pmc/articles/PMC8075504/

140. PHE-CRUK. Main Cancer Treatments: CAS-SOP #4.7 [Internet]. Available from: http://www.ncin.org.uk/cancer_type_and_topic_specific_work/topic_specific_work/main_cancer_treatments

141. Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick NK, Fatemifar G, Banerjee A, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. Journal of the American Medical Informatics Association. 2019 Nov 15;26(12):1545–59. Available from: https://www.caliberresearch.org

142. Payne RA, Mendonca SC, Elliott MN, Saunders CL, Edwards DA, Marshall M, et al. Development and validation of the Cambridge Multimorbidity Score. CMAJ. 2020 Feb 3;192(5):E107–14. Available from: https://www.cmaj.ca/content/192/5/E107

143. Moore SF, Price SJ, Chowienczyk S, Bostock J, Hamilton W. The impact of changing risk thresholds on the number of people in England eligible for urgent investigation for possible cancer: an observational cross-sectional study. Br J Cancer. 2021 Nov 23;125(11):1593–7. Available from: https://www.nature.com/articles/s41416-021-01541-4

144. Medicines & Healthcare products Regulatory Agency (MHRA). CPRD small area level data based on patient postcode: Documentation and Data Dictionary (Set 16). 2018.

145. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, Staa T van, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015;44(3):827–36.

146. Clinical Practice Research Datalink. CPRD GOLD February 2022 (Version 2022.02.001) [Data set] [Internet]. 2022. Available from: https://doi.org/10.48329/kxsp-t315

147. Green LA, Fryer GE, Yawn BP, Lanier D, Dovey SM. The Ecology of Medical Care Revisited. New England Journal of Medicine. 2001 Jun 28;344(26):2021–5. Available from: www.nejm.org

148. Nicholson BD, Mant D, Bankhead C. Can safety-netting improve cancer detection in patients with vague symptoms? BMJ (Online). 2016 Nov 9;355. Available from: https://www.bmj.com/content/355/bmj.i5515

149. Siminoff LA, Rogers HL, Harris-Haywood S. Missed opportunities for the diagnosis of colorectal cancer. Biomed Res Int. 2015;2015.

150. (NCRAS) NCR and AS. Routes to diagnosis 2006 to 2015: Technical document [Internet]. 2018. Available from: https://www.cancerdata.nhs.uk/routestodiagnosis/Routes_to_Diagnosis_2006_2015_technical_document.pdf

151. Office for National Statistics (ONS). Cancer Registration Statistics, England, 2011 [Internet]. 2011. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/cancerregistrationstatisticscancerregistrationstatisticsengland

152. Office for National Statistics (ONS). Mid-2011 Population Estimates: Single year of age and sex for local authorities in the United Kingdom [Internet]. 2013. Available from: https://webarchive.nationalarchives.gov.uk/20160107185425/http://www.ons.gov.uk/ons/rel/pop-estimate/population-estimates-for-uk--england-and-wales--scotland-and-northern-ireland/mid-2011-and-mid-2012/rft---mid-2011-uk-population-estimates.zip

153. Shah ASV, Wood R, Gribben C, Caldwell D, Bishop J, Weir A, et al. Risk of hospital admission with coronavirus disease 2019 in healthcare workers and their households: nationwide linkage cohort study. BMJ. 2020;371:m3582.

154. Reulen RC, Guha J, Bright CJ, Henson KE, Feltbower RG, Hall M, et al. Risk of cerebrovascular disease among 13,457 five-year survivors of childhood cancer: a population based cohort study. Int J Cancer. 2020; Available from: http://www.ncbi.nlm.nih.gov/pubmed/32683688

155. Barclay ME, Lyratzopoulos G, Walter FM, Jefferies S, Peake MD, Rintoul RC. Incidence of second and higher order smoking-related primary cancers following lung cancer: A population-based cohort study. Thorax. 2019 May 1;74(5):466–72. Available from: http://dx.doi.org/10.1136/thoraxjnl-2018-212456

156. Consonni D, Coviello E, Buzzoni C, Mensi C. A command to calculate age-standardized rates with efficient interval estimation. Stata Journal. 2012 Dec 1;12(4):688–701. Available from: https://journals.sagepub.com/doi/10.1177/1536867X1201200408?icid=int.sj-abstract.similar-articles.2

157. Dobson AJ, Kuulasmaa K, Eberle E, Scherer J. Confidence intervals for weighted sums of poisson parameters. Stat Med. 1991 Mar 1;10(3):457–62. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/sim.4780100317

158. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008 Apr;61(4):344–9. Available from: https://pubmed.ncbi.nlm.nih.gov/18313558/

159. Tyczynski J, Démaret E, Parkin D. Standards and Guidelines for Cancer Registration in Europe: The ENCR Recommendations [Internet]. IARC Technical Publication No. 40. International Agency for Research on Cancer (IARC); 2003. 1 p. Available from: https://publications.iarc.fr/Book-And-Report-Series/Iarc-Technical-Publications/Standards-And-Guidelines-For-Cancer-Registration-In-Europe-2003

160. Moore SF, Price SJ, Chowienczyk S, Bostock J, Hamilton W. The impact of changing risk thresholds on the number of people in England eligible for urgent investigation for possible cancer: an observational cross-sectional study. Br J Cancer. 2021 Nov 23;125(11):1593–7. Available from: /pmc/articles/PMC8445014/

161. Bhise V, Meyer AND, Menon S, Singhal G, Street RL, Giardina TD, et al. Patient perspectives on how physicians communicate diagnostic uncertainty: An experimental vignette study. International Journal for Quality in Health Care. 2018 Feb 1;30(1):2–8. Available from: https://academic.oup.com/intqhc/article/30/1/2/4791877

162. Gallagher AM, Thomas JM, Hamilton WT, White PD. Incidence of fatigue symptoms and diagnoses presenting in UK primary care from 1990 to 2001. J R Soc Med. 2004 Dec;97(12):571–5. Available from: /pmc/articles/PMC1079668/

163. White B, Rafiq M, Gonzalez-Izquierdo A, Hamilton W, Price S, Lyratzopoulos G. Risk of cancer following primary care presentation with fatigue: a population-based cohort study of a quarter of a million patients. Br J Cancer. 2022 Jun 1;126(11):1627–36. Available from: https://www.nature.com/articles/s41416-022-01733-6

164. Herbert A, Rafiq M, Pham TM, Renzi C, Abel GA, Price S, et al. Predictive values for different cancers and inflammatory bowel disease of 6 common abdominal symptoms among more than 1.9 million primary care patients in the UK: A cohort study. Basu S, editor. PLoS Med. 2021 Aug 2;18(8):e1003708. Available from: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003708

165. Chapman D, Poirier V, Vulkan D, Fitzgerald K, Rubin G, Hamilton W, et al. First results from five multidisciplinary diagnostic centre (MDC) projects for non-specific but concerning symptoms, possibly indicative of cancer. Br J Cancer. 2020 Sep 1;123(5):722–9. Available from: https://www.nature.com/articles/s41416-020-0947-y

166. Chapman D, Poirier V, Fitzgerald K, Nicholson BD, Hamilton W. Non-specific symptoms-based pathways for diagnosing less common cancers in primary care: a service evaluation. British Journal of General Practice. 2021 Nov 1;71(712):e846–53. Available from: https://bjgp.org/content/71/712/e846

167. Bouras G, Markar SR, Burns EM, Huddy JR, Bottle A, Athanasiou T, et al. The psychological impact of symptoms related to esophagogastric cancer resection presenting in primary care: A national linked database study. European Journal of Surgical Oncology (EJSO). 2017 Feb 1;43(2):454–60. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0748798316309489

168. Doran T, Kontopantelis E, Valderas JM, Campbell S, Roland M, Salisbury C, et al. Effect of financial incentives on incentivised and non-incentivised clinical activities: longitudinal analysis of data from the UK Quality and Outcomes Framework. BMJ. 2011 Jun 28;342(7814). Available from: https://www.bmj.com/content/342/bmj.d3590

169. Hawker JI, Smith S, Smith GE, Morbey R, Johnson AP, Fleming DM, et al. Trends in antibiotic prescribing in primary care for clinical syndromes subject to national recommendations to reduce antibiotic resistance, UK 1995-2011: analysis of a large database of primary care consultations. J Antimicrob Chemother. 2014 Dec 1;69(12):3423–30. Available from: https://pubmed.ncbi.nlm.nih.gov/25091508/

170. Kuan V, Denaxas S, Gonzalez-Izquierdo A, Direk K, Bhatti O, Husain S, et al. A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. Lancet Digit Health. 2019 Jun 1;1(2):e63–77. Available from: www.thelancet.com/

171. Masefield SC, Prady SL, Sheldon TA, Small N, Jarvis S, Pickett KE. The Effects of Caring for Young Children with Developmental Disabilities on Mothers' Health and Healthcare Use: Analysis of Primary Care Data in the Born in Bradford Cohort. J Dev Phys Disabil. 2021 Apr 8;34:67–87. Available from: https://link.springer.com/article/10.1007/s10882-021-09789-7

172. Palin V, Mölter A, Belmonte M, Ashcroft DM, White A, Welfare W, et al. Antibiotic prescribing for common infections in UK general practice: variability and drivers. Journal of Antimicrobial Chemotherapy. 2019 Aug 1;74(8):2440–50. Available from: https://academic.oup.com/jac/article/74/8/2440/5481890

173. Denaxes Lab. CALIBER phenotype portal: Haemoglobin measurement [Internet]. Available from: https://www.caliberresearch.org/portal/show/haemoglobin_gprd

174. Hamilton W, Lancashire R, Sharp D, Peters TJ, Cheng KK, Marshall T. The importance of anaemia in diagnosing colorectal cancer: A case-control study using electronic primary care records. Br J Cancer. 2008 Jan 29;98(2):323–7. Available from: /pmc/articles/PMC2361444/

175. Hamilton W, Round A, Sharp D, Peters TJ. Clinical features of colorectal cancer before diagnosis: A population-based case-control study. Br J Cancer. 2005;93(4):399–405.

176. Hamilton W, Kernick D. Clinical features of primary brain tumours: A case-control study using electronic primary care records. British Journal of General Practice. 2007 Sep;57(542):695–9. Available from: /pmc/articles/PMC2151783/

177. Hamilton W, Peters TJ, Bankhead C, Sharp D. Risk of ovarian cancer in women with symptoms in primary care: population based case-control study. BMJ. 2009 Aug 25;339(aug25 2):b2998–b2998. Available from: http://www.bmj.com/

178. Hamilton W, Sharp DJ, Peters TJ, Round AP. Clinical features of prostate cancer before diagnosis: A population-based, case-control study. British Journal of General Practice. 2006 Oct;56(531):756–62. Available from: /pmc/articles/PMC1920715/

179. Hippisley-Cox J, Coupland C. Identifying patients with suspected colorectal cancer in primary care: Derivation and validation of an algorithm. British Journal of General Practice. 2012 Jan 1;62(594):e29–37. Available from: https://bjgp.org/content/62/594/e29

180. Hippisley-Cox J, Coupland C. Identifying patients with suspected gastro-oesophageal cancer in primary care: Derivation and validation of an algorithm. British Journal of General Practice. 2011 Nov 1;61(592):e707–14. Available from: https://bjgp.org/content/61/592/e707

181. Hippisley-Cox J, Coupland C. Identifying women with suspected ovarian cancer in primary care: Derivation and validation of algorithm. BMJ (Online). 2012 Jan 28;344(7841). Available from: http://www.bmj.com/

182. Hippisley-Cox J, Coupland C. Identifying patients with suspected pancreatic cancer in primary care: Derivation and validation of an algorithm. British Journal of General Practice. 2012 Jan;62(594). Available from: https://pubmed.ncbi.nlm.nih.gov/22520674/

183. Hippisley-Cox J, Coupland C. Identifying patients with suspected lung cancer in primary care: Derivation and validation of an algorithm. British Journal of General Practice. 2011 Nov 1;61(592):e715–23. Available from: https://bjgp.org/content/61/592/e715

184. Hippisley-Cox J, Coupland C. Identifying patients with suspected renal tract cancer in primary care: Derivation and validation of an algorithm. British Journal of General Practice. 2012 Apr 1;62(597):e251–60. Available from: https://bjgp.org/content/62/597/e251

185. Nicholson BD, Ordóñez-Mena JM, Lay-Flurrie S, Sheppard JP, Liyanage H, McGagh D, et al. Consultations for clinical features of possible cancer and associated urgent referrals before and during the COVID-19 pandemic: an observational cohort study from English primary care. British Journal of Cancer 2021 126:6. 2021 Dec 21;126(6):948–56. Available from: https://www.nature.com/articles/s41416-021-01666-6

186. Koch H, van Bokhoven MA, ter Riet G, van Alphen-Jager JMT, van der Weijden T, Dinant GJ, et al. Ordering blood tests for patients with unexplained fatigue in general practice: what does it yield? Results of the VAMPIRE trial. Br J Gen Pract. 2009 Apr;59(561):243–9. Available from: https://pubmed.ncbi.nlm.nih.gov/19341544/

187. Watson J, de Salis I, Banks J, Salisbury C. What do tests do for doctors? A qualitative study of blood testing in UK primary care. Fam Pract. 2017 Nov 16;34(6):735–9. Available from: https://academic.oup.com/fampra/article/34/6/735/3867385

188. NHS England. National Cancer Waiting Times Monitoring Data Set: Two week wait cancer or symptomatic breast referral type [Internet]. Available from: https://www.datadictionary.nhs.uk/data_elements/two_week_wait_cancer_or_symptomatic_breast_referral_type.html

189. Wiering B, Lyratzopoulos G, Hamilton W, Campbell J, Abel G. Concordance with urgent referral guidelines in patients presenting with any of six 'alarm' features of possible cancer: a retrospective cohort study using linked primary care records. BMJ Qual Saf. 2022 Aug 1;31(8):579–89. Available from: https://qualitysafety.bmj.com/content/31/8/579

190. White B, Renzi C, Barclay M, Lyratzopoulos G. Underlying cancer risk among patients with fatigue and other vague symptoms: a population-based cohort study in primary care. British Journal of General Practice. 2023 Jan 9;BJGP.2022.0371. Available from: https://bjgp.org/content/early/2023/01/09/BJGP.2022.0371

191. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). Int J Epidemiol. 2017 Aug 1;46(4):1093–1093i. Available from: https://academic.oup.com/ije/article/46/4/1093/3072145

192. Medicines & Healthcare products Regulatory Agency (MHRA). Hospital Episode Statistics (HES) Admitted Patient Care and CPRD primary care data Documentation (Set 16). 2018.

193. Medicines & Healthcare products Regulatory Agency (MHRA). The Public Health England National Cancer Registration and Analysis Service (NCRAS) and CPRD primary care data Documentation (Set 16). 2018.

194. Hoile R, Tabet N, Smith H, Bremner S, Cassell J, Ford E. Are symptoms of insomnia in primary care associated with subsequent onset of dementia? A matched retrospective case-control study. Aging Ment Health. 2020 Sep 1;24(9):1466–71. Available from: https://pubmed.ncbi.nlm.nih.gov/31791142/

195. Cairns V, Wallenhorst C, Rietbrock S, Martinez C. Incidence of Lyme disease in the UK: a population-based cohort study. BMJ Open. 2019 Jul 1;9(7):e025916. Available from: https://bmjopen.bmj.com/content/9/7/e025916

196. Whitfield E. Atlas-phenotypes: Codelists of Read v2, Snomed and ICD10 codes for conditions included in Atlas collated from multiple sources and code browsers. [Internet]. Available from: https://github.com/ekw26/Atlas-phenotypes

197. Ramirez A. Pneumonia- VUMC eMERGE v5.1 | PheKB [Internet]. Available from: https://phekb.org/phenotype/pneumonia-vumc-emerge-v51

198. Kontopantelis E, Olier I, Planner C, Reeves D, Ashcroft DM, Gask L, et al. Primary care consultation rates among people with and without severe mental illness: A UK cohort study using the Clinical Practice Research Datalink. BMJ Open. 2015 Dec 1;5(12):e008650. Available from: https://bmjopen.bmj.com/content/5/12/e008650

199. Whitfield E;, White B;, Denaxas S;, Lyratzopoulos, Georgios. Diagnostic windows in non-neoplastic diseases: A systematic review. British Journal of General Practice. 2023 Apr 12;BJGP.2023.0044. Available from: https://bjgp.org/content/early/2023/04/04/BJGP.2023.0044

200. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999 Jan;10(1):37–48. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9888278

201. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ. 2017 May 23;357:j2099. Available from: https://www.bmj.com/content/357/bmj.j2099

202. National Institute for Health and Care Excellence. Vol. 49, NICE guideline NG136. 2019. Hypertension in adults: Diagnosis and management. [A] Evidence review for diagnosis. Available from: https://www.nice.org.uk/guidance/ng136/evidence/a-diagnosis-pdf-6896748206

203. Department of Health. Direct Access to Diagnostic Tests for Cancer: Best Practice Referral Pathways for General Practitioners. 2011; Available from: www.nationalarchives.gov.uk/doc/open-government-licence/

204. Kim D, Aylin P, Bottle A, Hayhoe B, Majeed A, Cowie MR. Route to heart failure diagnosis in English primary care: a retrospective cohort study of variation. British Journal of General Practice. 2019 Oct 1;69(687):e697–705. Available from: https://bjgp.org/content/69/687/e697

205. Sandler CX, Wyller VBB, Moss-Morris R, Buchwald D, Crawley E, Hautvast J, et al. Long COVID and Post-infective Fatigue Syndrome: A Review. Open Forum Infect Dis. 2021 Oct 1;8(10). Available from: /pmc/articles/PMC8496765/

206. Medina-Lara A, Grigore B, Lewis R, Peters J, Price S, Landa P, et al. Cancer diagnostic tools to aid decision-making in primary care: Mixed-methods systematic reviews and cost-effectiveness analysis. Health Technol Assess (Rockv). 2020 Nov 1;24(66):1–366.

207. Paige E, Barrett J, Stevens D, Keogh RH, Sweeting MJ, Nazareth I, et al. Practice of Epidemiology Landmark Models for Optimizing the Use of Repeated Measurements of Risk Factors in Electronic Health Records to Predict Future Disease Risk. Am J Epidemiol. 2018;187(7):1530–8. Available from: http://creativecommons.

208. Rizopoulos D, Molenberghs G, Lesaffre EMEH. Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. Biometrical Journal. 2017 Nov 9;59(6):1261–76. Available from: https://onlinelibrary.wiley.com/doi/10.1002/bimj.201600238

209. Li W, Li L, Astor BC. A comparison of two approaches to dynamic prediction: Joint modeling and landmark modeling. Stat Med. 2023 Jun 15;42(13):2101–15. Available from: https://onlinelibrary.wiley.com/doi/10.1002/sim.9713

210. Placido D, Yuan B, Hjaltelin JX, Zheng C, Haue AD, Chmura PJ, et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. Nat Med. 2023 May 8;29(5):1113–22. Available from: https://doi.org/10.1038/s41591-023-02332-5

211. Whitfield E, White B, Denaxas S, Lyratzopoulos G. Diagnostic windows in non-neoplastic diseases: A systematic review. British Journal of General Practice. 2023 Apr 12;BJGP.2023.0044. Available from: https://bjgp.org/content/early/2023/04/04/BJGP.2023.0044

212. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLoS Med. 2013;10(2):e1001381. Available from: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001381

# 10. Appendices

## 10.2 Chapter 2 appendices

### 10.2.1 UCL Research paper declaration form

# UCL Research Paper Declaration Form

## referencing the doctoral candidate's own published work(s)

*Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:*

- *have been uploaded to a preprint server*
- *are in submission to a peer-reviewed publication*
- *have been published in a peer-reviewed publication, e.g. journal, textbook.*

*This form should be completed as many times as necessary. For instance, if you have seven thesis chapters, two of which containing material that has already been published, you would complete this form twice.*

1. **For a research manuscript that has already been published** (if not yet published, please skip to section 2)

   a) **What is the title of the manuscript?**

   Does changing healthcare use signal opportunities for earlier detection of cancer? A review of studies using information from electronic patient records.

   b) **Please include a link to or doi for the work**

   https://doi.org/10.1016/j.canep.2021.102072

   c) **Where was the work published?**

   Cancer Epidemiology

   d) **Who published the work?** (e.g. OUP)

   Elsevier

   e) **When was the work published?**

   2022

   f) **List the manuscript's authors in the order they appear on the publication**

   White, B., Renzi, C., Rafiq, M., Abel, G. A., Jensen, H., & Lyratzopoulos, G.

   g) **Was the work peer reviewed?**

   Yes

   h) **Have you retained the copyright?**

   Yes

   i) **Was an earlier form of the manuscript uploaded to a preprint server?** (e.g. medRxiv). If 'Yes', please give a link or doi)

No

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

☐

*I acknowledge permission of the publisher named under **1d** to include in this thesis portions of the publication named as included in **1c**.*

**2. For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)

BW, GL, and CR conceived and designed the study and agreed the search and data extraction strategy. BW identified and analysed the studies, under the supervision of GL and CR and with a sample of studies independently reviewed by CR. MR, GL and CR provided clinical input into interpretations. GA and HJ provided statistical and methodological expertise. All authors contributed to drafting and revising the article.

**3. In which chapter(s) of your thesis can this material be found?**

**2**

**4. e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

*Candidate*

*Date:*

05/06/2023

*Supervisor/ Senior Author (where appropriate)*

*Date*

*13/06/2023*

Click or tap here to enter text.

### 10.2.2 Pubmed search terms for author search

cancer[Filter] AND (early detection of cancer[MeSH Terms] OR signs and symptoms[MeSH Terms] OR "before diagnosis" OR pre-diagnos* OR prediagnos*) AND (Ahrensberg JM[Author] OR Christensen KG [Author] OR Chu TPC[Author] OR "Ewing M"[Author] OR Friis Abrahamsen C[Author] OR Guldbrandt LM[Author] OR Hansen PL[Author] OR Hauswaldt J[Author] OR Koshiaris C[Author] OR "Jensen H"[Author] OR "McDonald L"[Author] OR Morrell S[Author] OR Nygaard C[Author] OR Pottegard A[Author] OR Raedkjaer M[Author] OR Renzi C[Author] OR "Wang Yingying"[Author] OR "Zhou Yin"[Author] OR "Nanna Holt Jessen"[Author] OR "Kuiper JG"[Author])

## 10.3 Chapter 3 appendices

### 10.3.1 Inclusion criteria to select patients for the pre-selected data included in CPRD extract #1

The following is the inclusion criteria agreed with CPRD in the study data specification.

**Pre-selected symptom cohorts**

From the source population in CPRD GOLD:

- 16,343,492 patients were acceptable
- 8,041,292 patients were eligible for linkage to the CR cancer registry database
- 3,342,277 patients had a record for at least one of the selected symptoms (Abdominal pain; Abdominal bloating; Breast lump; Change in bowel habit; Dyspepsia; Dysphagia; Dyspnoea/shortness of breath; Fatigue; Haematuria; Haemoptysis; Jaundice; Night sweats; Post-menopausal bleeding; Rectal bleeding; Weight loss) in their clinical or referral file based on the Read codes received.
- 3,342,257 patients met gender criterion (Male and Female Only)
- 1,168,842 patients had at least one event record:
    - within the study period (01/01/2007 – 31/12/2016)
    - within their up-to-standard (UTS) period.
    - while aged 30-99 years.
    - with at least one year of UTS follow-up in CPRD GOLD before the event date.

**1,168,842** patients were included in the study population.

*Clinical Practice Research Datalink **National Cancer Registry and Analysis Service

### 10.3.2 Inclusion criteria to select patients for the pre-selected data included in CPRD extract #2

The following is the inclusion criteria agreed with CPRD in the study data specification.

**Pre-selected symptom cohorts**

From the source population in CPRD GOLD:

- 20,620,714 patients were acceptable
- 20,619,636 patients met gender criterion (Male and Female Only)
- 8,399,126 patients were eligible for linkage to the CR cancer registry database
- 5,985,788 patients had a record for one of the selected features:
    - Symptoms (Abdominal pain; Abdominal bloating; Breast lump; Change in bowel habit; Dyspepsia; Dysphagia; Dyspnoea/shortness of breath; Fatigue; Haematuria; Haemoptysis; Jaundice; Night sweats; Post-menopausal bleeding; Rectal bleeding; Weight loss; abdominal lump/mass; constipation; cough; diarrhoea; pelvic pain; stomach disorders; urinary tract infections) recorded in their clinical, test, or referral file based on the Read codes received;
    - blood test-related features (full blood count, haemoglobin concentration; platelet count; albumin; CRP- C reactive protein; ESR- Erythrocyte sedimentation rate; PV- Plasma viscosity; ferritin) in their clinical, test or referral file based on the Read codes received, or in their test file based on the entity types received
- 2,530,253 patients had at least one event record:
    - within the study period (01/01/2007 – – 31/10/2021.
    - within their up-to-standard (UTS) period.
    - while aged 30-99 years.

**2,530,253 patients** were included in the study population.

**Random sample**

The control (reference) group will be a random sample of **one million patients** registered with CPRD during the study period 01/01/2007- 31/10/2021 meeting the following inclusion criteria:

- 'Acceptable' patients
- Patients should be eligible for linkage to the CR cancer registry database.
- Only ***male/ female*** patients should be included
- had at least one year of UTS follow up between 01/01/2007 to 31/10/2021, while they were also aged 30-99 years

No further restrictions were applied to the reference group.

*Clinical Practice Research Datalink **National Cancer Registry and Analysis Service

### 10.3.3 List of Read codes used to define fatigue in CPRD

| medcode | readcode | readcode_desc |
|---|---|---|
| 1042 | R007400 | [D]Postviral (asthenic) syndrome |
| 1147 | R007500 | [D]Tiredness |
| 1371 | R007300 | [D]Lethargy |
| 1404 | 1682.00 | Fatigue |
| 1582 | E205.11 | Nervous exhaustion |
| 1688 | R007100 | [D]Fatigue |
| 1900 | R2y3.00 | "[D]Debility, unspecified" |
| 2855 | 1B32.00 | Weakness present |
| 3361 | E205.00 | Neurasthenia - nervous debility |
| 4546 | F286.00 | Chronic fatigue syndrome |
| 5049 | R007200 | [D]Asthenia NOS |
| 5583 | 168..12 | Lethargy - symptom |
| 5658 | R007000 | [D]Malaise |
| 5751 | 1683.00 | Tired all the time |
| 5794 | 168..00 | Tiredness symptom |
| 5814 | R007z11 | [D]Lassitude |
| 6029 | 1B3..12 | Weakness symptoms |
| 6190 | F286.12 | Postviral fatigue syndrome |
| 6242 | 168..11 | Fatigue - symptom |
| 7235 | E205.12 | Tired all the time |
| 7529 | F286.11 | CFS - Chronic fatigue syndrome |
| 9127 | F286.14 | Post-viral fatigue syndrome |
| 9220 | 1688.00 | Exhaustion |
| 9435 | 2254.00 | O/E - apathetic |
| 9656 | Eu46011 | [X]Fatigue syndrome |
| 9823 | 1684.11 | C/O - debility - malaise |
| 9889 | R007211 | [D]General weakness |
| 12411 | R007411 | [D]Post viral debility |
| 15516 | 1683.11 | C/O - 'tired all the time' |
| 16479 | 2832.12 | O/E - weakness |
| 16561 | Eu46000 | [X]Neurasthenia |
| 17736 | 1684.00 | Malaise/lethargy |
| 23932 | R007z00 | [D]Malaise and fatigue NOS |
| 24382 | R204.00 | [D]Senile exhaustion |
| 27877 | F286.13 | PVFS - Postviral fatigue syn |
| 29292 | 168Z.00 | Tiredness symptom NOS |
| 44215 | R007.00 | [D]Malaise and fatigue |
| 97284 | F286100 | Moderate chronic fatigue syndrome |
| 98512 | F286000 | Mild chronic fatigue syndrome |
| 98734 | F286200 | Severe chronic fatigue syndrome |

## 10.4 Chapter 4 appendices

### 10.4.1 UCL Research paper declaration form

# UCL Research Paper Declaration Form
## referencing the doctoral candidate's own published work(s)

*Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:*

- *have been uploaded to a preprint server*
- *are in submission to a peer-reviewed publication*
- *have been published in a peer-reviewed publication, e.g. journal, textbook.*

*This form should be completed as many times as necessary. For instance, if you have seven thesis chapters, two of which containing material that has already been published, you would complete this form twice.*

1. **For a research manuscript that has already been published** (if not yet published, please skip to section 2)

   a) **What is the title of the manuscript?**

   Risk of cancer following primary care presentation with fatigue: a population-based cohort study of a quarter of a million patients.

   b) **Please include a link to or doi for the work**

   https://doi.org/10.1038/s41416-022-01733-6

   c) **Where was the work published?**

   British Journal of Cancer

   d) **Who published the work?** (e.g. OUP)

   Nature publishing group

   e) **When was the work published?**

   2022

   f) **List the manuscript's authors in the order they appear on the publication**

   White, B., Rafiq, M., Gonzalez-Izquierdo, A., Hamilton, W., Price, S., & Lyratzopoulos, G.

   g) **Was the work peer reviewed?**

   Yes

   h) **Have you retained the copyright?**

   Yes

i) **Was an earlier form of the manuscript uploaded to a preprint server?** (e.g. medRxiv). If 'Yes', please give a link or doi)

No

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

☐

*I acknowledge permission of the publisher named under **1d** to include in this thesis portions of the publication named as included in **1c**.*

**2. For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)

BW, GL, and MR conceived and designed the study. BW and AGI managed and BW analysed the data, under the supervision of GL, MR, and AGI. MR and GL provided clinical input, and WH and SP developed medical code lists used for case identification and advised on the presentation and discussion of results. All authors contributed to drafting and revising the article.

**3. In which chapter(s) of your thesis can this material be found?**

**4**

**4. e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

*Candidate*

*Date:*

05/06/2023

*Date*

*13/06/2023*

Click or tap here to enter text.

### 10.4.2 Read codes used to define fatigue

| CPRD Medcode | Readcode | Read term |
|---|---|---|
| 5794 | 168..00 | Tiredness symptom |
| 5751 | 1683 | Tired all the time |
| 1147 | R007500 | [D]Tiredness |
| 15516 | 1683.11 | C/O - 'tired all the time' |
| 5583 | 168..12 | Lethargy - symptom |
| 1404 | 1682 | Fatigue |
| 6029 | 1B3..12 | Weakness symptoms |
| 7235 | E205.12 | Tired all the time |
| 6242 | 168..11 | Fatigue - symptom |
| 5658 | R007000 | [D]Malaise |
| 1371 | R007300 | [D]Lethargy |
| 7529 | F286.11 | CFS - Chronic fatigue syndrome |
| 1688 | R007100 | [D]Fatigue |
| 4546 | F286.00 | Chronic fatigue syndrome |
| 17736 | 1684 | Malaise/lethargy |
| 1900 | R2y3.00 | [D]Debility, unspecified |
| 9220 | 1688 | Exhaustion |
| 2855 | 1B32.00 | Weakness present |
| 5814 | R007z11 | [D]Lassitude |
| 9823 | 1684.11 | C/O - debility - malaise |
| 29292 | 168Z.00 | Tiredness symptom NOS |
| 6190 | F286.12 | Postviral fatigue syndrome |
| 9889 | R007211 | [D]General weakness |
| 1582 | E205.11 | Nervous exhaustion |
| 5049 | R007200 | [D]Asthenia NOS |
| 1042 | R007400 | [D]Postviral (asthenic) syndrome |
| 16479 | 2832.12 | O/E - weakness |
| 3361 | E205.00 | Neurasthenia - nervous debility |
| 9127 | F286.14 | Post-viral fatigue syndrome |
| 9656 | Eu46011 | [X]Fatigue syndrome |
| 23932 | R007z00 | [D]Malaise and fatigue NOS |
| 9435 | 2254 | O/E - apathetic |
| 12411 | R007411 | [D]Post viral debility |
| 27877 | F286.13 | PVFS - Postviral fatigue syn |
| 16561 | Eu46000 | [X]Neurasthenia |
| 44215 | R007.00 | [D]Malaise and fatigue |
| 24382 | R204.00 | [D]Senile exhaustion |
| 97284 | F286100 | Moderate chronic fatigue syndrome |
| 98512 | F286000 | Mild chronic fatigue syndrome |
| 98734 | F286200 | Severe chronic fatigue syndrome |

## 10.4.3 Number and cumulative proportion of patients with fatigue diagnosed with cancer, by month of first cancer diagnosis, observed compared to expected

| | Observed[a] | | | Expected[b] | | | Excess cases | | Cumulative excess cases | |
|---|---|---|---|---|---|---|---|---|---|---|
| Months | Cases (n) | Cumulative cases (n) | Cumulative cases per 1,000 patients [%,lci,uci] | Cases (n) | Cumulative cases (n) | Cumulative cases per 1,000 patients [%,lci,uci] | n | P-value | n | Cases per 1,000 total patients |
| 1 | 856 | 856 | 3.4 [3.2,3.7] | 194 | 194 | 0.8 [0.8,0.8] | 662 | <0.001 | 662 | 2.6 |
| 2 | 649 | 1,505 | 6 [5.7,6.3] | 194 | 388 | 1.5 [1.5,1.6] | 455 | <0.001 | 1117 | 4.5 |
| 3 | 411 | 1,916 | 7.6 [7.3,8] | 194 | 582 | 2.3 [2.3,2.3] | 217 | <0.001 | 1334 | 5.3 |
| 4 | 331 | 2,247 | 9 [8.6,9.3] | 194 | 776 | 3.1 [3.1,3.1] | 137 | <0.001 | 1471 | 5.9 |
| 5 | 293 | 2,540 | 10.1 [9.7,10.5] | 194 | 970 | 3.9 [3.8,3.9] | 99 | <0.001 | 1570 | 6.3 |
| 6 | 276 | 2,816 | 11.2 [10.8,11.7] | 194 | 1,164 | 4.6 [4.6,4.7] | 82 | <0.001 | 1652 | 6.6 |
| 7 | 209 | 3,025 | 12.1 [11.6,12.5] | 194 | 1,358 | 5.4 [5.4,5.4] | 15 | 0.285 | 1667 | 6.7 |
| 8 | 250 | 3,275 | 13.1 [12.6,13.5] | 194 | 1,552 | 6.2 [6.2,6.2] | 56 | <0.001 | 1723 | 6.9 |
| 9 | 222 | 3,497 | 14 [13.5,14.4] | 194 | 1,746 | 7 [6.9,7] | 28 | 0.049 | 1751 | 7.0 |
| 10 | 198 | 3,695 | 14.7 [14.3,15.2] | 194 | 1,940 | 7.7 [7.7,7.8] | 4 | 0.767 | 1755 | 7.0 |
| 11 | 196 | 3,891 | 15.5 [15,16] | 194 | 2,134 | 8.5 [8.5,8.6] | 2 | 0.877 | 1757 | 7.0 |
| 12 | 196 | 4,087 | 16.3 [15.8,16.8] | 194 | 2,328 | 9.3 [9.3,9.3] | 2 | 0.877 | 1759 | 7.0 |
| 13 | 181 | 4,268 | 17 [16.5,17.5] | 194 | 2,522 | 10.1 [10,10.1] | -13 | 0.353 | 1746 | 7.0 |
| 14 | 191 | 4,459 | 17.8 [17.3,18.3] | 194 | 2,716 | 10.8 [10.8,10.9] | -3 | 0.839 | 1743 | 7.0 |
| 15 | 201 | 4,660 | 18.6 [18.1,19.1] | 194 | 2,910 | 11.6 [11.6,11.7] | 7 | 0.611 | 1750 | 7.0 |
| 16 | 216 | 4,876 | 19.5 [18.9,20] | 194 | 3,104 | 12.4 [12.3,12.4] | 22 | 0.120 | 1772 | 7.1 |
| 17 | 194 | 5,070 | 20.2 [19.7,20.8] | 194 | 3,298 | 13.2 [13.1,13.2] | 0 | 1.009 | 1772 | 7.1 |
| 18 | 184 | 5,254 | 21 [20.4,21.5] | 194 | 3,492 | 13.9 [13.9,14] | -10 | 0.478 | 1762 | 7.0 |
| 19 | 184 | 5,438 | 21.7 [21.1,22.3] | 194 | 3,686 | 14.7 [14.7,14.8] | -10 | 0.478 | 1752 | 7.0 |
| 20 | 178 | 5,616 | 22.4 [21.8,23] | 194 | 3,880 | 15.5 [15.4,15.5] | -16 | 0.251 | 1736 | 6.9 |
| 21 | 180 | 5,796 | 23.1 [22.5,23.7] | 194 | 4,074 | 16.3 [16.2,16.3] | -14 | 0.317 | 1722 | 6.9 |
| 22 | 158 | 5,954 | 23.8 [23.2,24.4] | 194 | 4,268 | 17 [17,17.1] | -36 | 0.008 | 1686 | 6.7 |
| 23 | 177 | 6,131 | 24.5 [23.9,25.1] | 194 | 4,462 | 17.8 [17.8,17.9] | -17 | 0.222 | 1669 | 6.7 |
| 24 | 197 | 6,328 | 25.3 [24.6,25.9] | 194 | 4,656 | 18.6 [18.5,18.6] | 3 | 0.822 | 1672 | 6.7 |
| Total patients | | 250,606 | | | 250,606 | | | | | |

[a]Cancer diagnoses between 2007-2015, up to 24 months after first presentation with fatigue to primary care in 2007-2013. [b]Expected cases for the age/ sex distribution of patients with fatigue, based on five-year age band and sex-specific estimated monthly population incidence, using annual number of cancer diagnoses and mid-year population estimates for England, 2011.

### 10.4.4 International Classification of Diseases (ICD)-10 codes used to define all cancers combined, and each cancer site

| Cancer site | ICD10 codes |
|---|---|
| All malignant cancers excl. non melanoma skin cancer (including selected non-malignant cancers as specified) | C00-C97 (excluding C44) |
| Bladder | C67 |
| Malignant brain and other CNS | C70-72 |
| Breast | C50 |
| Cancer of unknown primary | C77-80 |
| Cervix | C53 |
| Colorectal | C18, C19, C20 |
| Head & neck | C00-C14, C31, C32 |
| Lymphoma (non-hodgkins & hodgkins) | C81, C82-85 |
| Kidney | C64 |
| Leukaemia | C91-95 |
| Liver | C22 |
| Lung & mesothelioma | C33, C34, C45 |
| Melanoma | C43 |
| Multiple myeloma | C90 |
| Ovary | C56-57 |
| Pancreas | C25 |
| Prostate | C61 |
| Sarcoma (soft tissue, connective & bone) | C40-41, C48-49 |
| Testis | C62 |
| Thyroid | C73 |
| Upper gastro-intestinal | C15-C16 |
| Uterus | C54-55 |
| Vulva | C51 |

### 10.4.5 STROBE Statement—Checklist of items that should be included in reports of cohort studies

| | Item No | Recommendation | Section & paragraph number |
|---|---|---|---|
| **Title and abstract** | 1 | (*a*) Indicate the study's design with a commonly used term in the title or the abstract<br>(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found | Title/ abstract<br><br>Abstract |
| **Introduction** | | | |
| Background/ rationale | 2 | Explain the scientific background and rationale for the investigation being reported | Background para. 1-3 |
| Objectives | 3 | State specific objectives, including any prespecified hypotheses | Background para. 4 |
| **Methods** | | | |
| Study design | 4 | Present key elements of study design early in the paper | Methods: Study design and data source |
| Setting | 5 | Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection | Methods: Study design and data source |
| Participants | 6 | (*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up<br>(*b*) For matched studies, give matching criteria and number of exposed and unexposed | Methods: Cohort identification<br><br>N/a – no matching |
| Variables | 7 | Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable | Methods: Follow up and outcomes; Methods: Statistical analysis para. 1 |
| Data sources/ measurement | 8* | For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group | Methods: Study design and data source |
| Bias | 9 | Describe any efforts to address potential sources of bias | Methods: Statistical analysis, para. 1-3 |
| Study size | 10 | Explain how the study size was arrived at | Methods: Statistical analysis para. 1 |
| Quantitative variables | 11 | Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why | Methods: Follow up and outcomes para. 2; Methods: Statistical analysis para. 1 |
| Statistical methods | 12 | (*a*) Describe all statistical methods, including those used to control for confounding<br>(*b*) Describe any methods used to examine subgroups and interactions<br>(*c*) Explain how missing data were addressed<br>(*d*) If applicable, explain how loss to follow-up was addressed<br>(*e*) Describe any sensitivity analyses | Methods: Statistical analysis<br><br>Methods: Statistical analysis para. 1<br>N/a – no missing<br>Methods: Follow up and outcomes para. 1<br>Methods: Cohort identification para. 3-4 |
| **Results** | | | |
| Participants | 13* | (a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed | Results: Cohort description |

| | | | |
|---|---|---|---|
| | | (b) Give reasons for non-participation at each stage<br>(c) Consider use of a flow diagram | Results: Cohort description<br>Results: Cohort description Fig. 1 |
| Descriptive data | 14* | (a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders<br>(b) Indicate number of participants with missing data for each variable of interest<br>(c) Summarise follow-up time (eg, average and total amount) | Results: Cohort description & Table 1.<br><br>N/a – no missing<br><br>N/a – complete follow up |
| Outcome data | 15* | Report numbers of outcome events or summary measures over time | Results: Risk of cancer & Table 2. |
| Main results | 16 | (*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included<br><br>(*b*) Report category boundaries when continuous variables were categorized<br>(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period | Findings: Risk of cancer; Findings: Frequency of specific cancer sites; Findings: Distribution of incident cases by month following recorded fatigue<br>N/a – no continuous variables<br><br>Findings: Table 2, Table 3, Fig. 2, Fig. 3 |
| Other analyses | 17 | Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses | Findings: Sensitivity analyses |
| **Discussion** | | | |
| Key results | 18 | Summarise key results with reference to study objectives | Discussion: Key findings |
| Limitations | 19 | Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias | Discussion: Strengths and limitations |
| Interpretation | 20 | Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence | Discussion: Implications |
| Generalisability | 21 | Discuss the generalisability (external validity) of the study results | Discussion: Strengths and limitations para. 3 |
| **Other information** | | | |
| Funding | 22 | Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based | Additional information: Funding information |

*Give information separately for exposed and unexposed groups.

**Note:** An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.

### 10.4.6 Deprivation quintile of patients presenting to primary care with fatigue, compared to England, by gender

| | Men | | | | Women | | | |
|---|---|---|---|---|---|---|---|---|
| | Patients with fatigue | | England population[b] | | Patients with fatigue | | England population[b] | |
| | N | % | N | % | N | % | N | % |
| Deprivation quintile[a] | | | | | | | | |
| 1 - least deprived | 18,961 | 23.37 | 2,861,959 | 17.90 | 39,462 | 23.29 | 3,039,573 | 17.72 |
| 2 | 18,639 | 22.97 | 3,117,552 | 19.50 | 37,675 | 22.23 | 3,299,259 | 19.23 |
| 3 | 17,604 | 21.70 | 3,301,569 | 20.65 | 36,178 | 21.35 | 3,539,247 | 20.63 |
| 4 | 14,084 | 17.36 | 3,355,807 | 20.99 | 30,086 | 17.75 | 3,627,502 | 21.14 |
| 5 - most deprived | 11,793 | 14.53 | 3,349,259 | 20.95 | 25,966 | 15.32 | 3,651,392 | 21.28 |
| Missing | 62 | 0.08 | - | - | 96 | 0.06 | - | - |
| Total people | 81,143 | | 15,986,146 | | 169,463 | | 17,156,973 | |

[a]Index of Multiple Deprivation (IMD) quintile of the person's area of residence. [b]Published statistics for men/ women aged 30 years and over in England in 2011, available at: www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/adhocs/12386populationbyindexofmultipledeprivationimdengland2001to2019

## 10.4.7 Number and proportion of patients diagnosed with cancer within a year after presenting to primary care with fatigue, by gender and index of multiple deprivation

| | Men | | | Women | | |
|---|---|---|---|---|---|---|
| | Cancer[a] | | Total | Cancer[a] | | Total |
| | n | % [lci,uci] | N | n | % [lci,uci] | N |
| Deprivation quintile[b] | | | | | | |
| 1 - least deprived | 474 | 2.5 [2.28,2.74] | 18,961 | 465 | 1.18 [1.07,1.29] | 39,462 |
| 2 | 458 | 2.46 [2.24,2.69] | 18,639 | 491 | 1.3 [1.19,1.42] | 37,675 |
| 3 | 416 | 2.36 [2.14,2.6] | 17,604 | 465 | 1.29 [1.17,1.41] | 36,178 |
| 4 | 334 | 2.37 [2.12,2.64] | 14,084 | 370 | 1.23 [1.11,1.36] | 30,086 |
| 5 - most deprived | 302 | 2.56 [2.28,2.87] | 11,793 | 308 | 1.19 [1.06,1.33] | 25,966 |

[a]Cancer diagnoses between 2007-2014, 12 months after first presentation, for patients presenting to primary care with a valid fatigue symptom between 2007-2013. [b]Index of Multiple Deprivation (IMD) quintile of the person's area of residence. Data not shown for missing IMD to reduce statistical disclosure risk.

### 10.4.8 Records of specific fatigue read codes, including all chronic fatigue syndrome and post-viral fatigue syndrome codes, as a proportion of all eligible records of fatigue between 2007-2013

| Read code | Read code description | Number of records n | Proportion of total records of fatigue % |
|---|---|---|---|
| 168..00 | Tiredness symptom | 100,126 | 25.97 |
| 1683.00 | Tired all the time | 85,050 | 22.06 |
| R007500 | [D]Tiredness | 29,903 | 7.76 |
| 1683.11 | C/O - 'tired all the time' | 23,822 | 6.18 |
| 168..12 | Lethargy - symptom | 23,136 | 6.00 |
| 1B3..12 | Weakness symptoms | 20,220 | 5.24 |
| 1682.00 | Fatigue | 19,410 | 5.03 |
| E205.12 | Tired all the time | 18,298 | 4.75 |
| 168..11 | Fatigue - symptom | 17,016 | 4.41 |
| R007000 | [D]Malaise | 12,429 | 3.22 |
| R007300 | [D]Lethargy | 10,182 | 2.64 |
| *F286.11 | CFS - Chronic fatigue syndrome | 3,024 | 0.78 |
| R007100 | [D]Fatigue | 2,053 | 0.53 |
| *F286.00 | Chronic fatigue syndrome | 1,871 | 0.49 |
| 1684.00 | Malaise/lethargy | 1,642 | 0.43 |
| R2y3.00 | [D]Debility, unspecified | 1,527 | 0.40 |
| 1688.00 | Exhaustion | 1,489 | 0.39 |
| 1B32.00 | Weakness present | 1,265 | 0.33 |
| R007z11 | [D]Lassitude | 1,235 | 0.32 |
| 1684.11 | C/O - debility - malaise | 1,054 | 0.27 |
| 168Z.00 | Tiredness symptom NOS | 692 | 0.18 |
| **F286.12 | Postviral fatigue syndrome | 636 | 0.16 |
| R007211 | [D]General weakness | 567 | 0.15 |
| E205.11 | Nervous exhaustion | 548 | 0.14 |
| R007200 | [D]Asthenia NOS | 407 | 0.11 |
| **R007400 | [D]Postviral (asthenic) syndrome | 353 | 0.09 |
| E205.00 | Neurasthenia - nervous debility | 295 | 0.08 |
| 2832.12 | O/E - weakness | 271 | 0.07 |
| **F286.14 | Post-viral fatigue syndrome | 220 | 0.06 |
| Eu46011 | [X]Fatigue syndrome | 169 | 0.04 |
| R007z00 | [D]Malaise and fatigue NOS | 113 | 0.03 |
| 2254.00 | O/E - apathetic | 94 | 0.02 |
| **R007411 | [D]Post viral debility | 89 | 0.02 |
| **F286.13 | PVFS - Postviral fatigue syn | 62 | 0.02 |
| Eu46000 | [X]Neurasthenia | 19 | 0.00 |
| R204.00 | [D]Senile exhaustion | 12 | 0.00 |
| R007.00 | [D]Malaise and fatigue | 10 | 0.00 |
| *F286100 | Moderate chronic fatigue syndrome | <5[a] | - |
| *F286200 | Severe chronic fatigue syndrome | <5[a] | - |
| *F286000 | Mild chronic fatigue syndrome | <5[a] | - |
| | *All chronic fatigue syndrome codes | 4,895 | 1.27 |
| | **All post-viral fatigue syndrome codes | 1,360 | 0.35 |
| | All records of fatigue | | 385,564 |

[a]Cell counts under 5 are suppressed to reduce statistical disclosure risk.

### 10.4.9 Number and proportion of patients whose index fatigue presentation was CFS or PVFS, by gender

| | Men | | Women | |
|---|---|---|---|---|
| | n | % | n | % |
| Chronic fatigue syndrome (CFS) only | 356 | 0.44 | 903 | 0.53 |
| Post-viral fatigue syndrome (PVFS) only | 240 | 0.30 | 534 | 0.32 |
| CFS or PVFS | 596 | 0.73 | 1,437 | 0.85 |
| All patients with fatigue including CFS & PVFS | 81,143 | | 169,463 | |

[a]Patients presenting to primary care with a valid fatigue symptom between 2007-2013.

### 10.4.10 Number and proportion of patients diagnosed with cancer within a year after presenting to primary care with fatigue, excluding patients whose index fatigue presentation was CFS or PVFS, by gender

| | Cancer[a] | | Total patients |
|---|---|---|---|
| | n | % | N |
| Chronic fatigue syndrome (CFS) only | 6 | 0.48 | 1,259 |
| Post-viral fatigue syndrome (PVFS) only | 5 | 0.65 | 774 |
| All fatigue codes excluding CFS & PVFS | 4,076 | 1.64 | 248,573 |
| All fatigue codes including CFS & PVFS | 4,087 | 1.63 | 250,606 |

[a]Cancer diagnoses between 2007-2014, 12 months after first presentation, for patients presenting to primary care with a valid fatigue symptom between 2007-2013. Results shown for men and women combined to reduce statistical disclosure risk.

## 10.5 Chapter 5 appendices

### 10.5.1 UCL Research paper declaration form

# UCL Research Paper Declaration Form
## referencing the doctoral candidate's own published work(s)

*Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:*

- *have been uploaded to a preprint server*
- *are in submission to a peer-reviewed publication*
- *have been published in a peer-reviewed publication, e.g. journal, textbook.*

*This form should be completed as many times as necessary. For instance, if you have seven thesis chapters, two of which containing material that has already been published, you would complete this form twice.*

1. **For a research manuscript that has already been published** (if not yet published, please skip to section 2)

   a) **What is the title of the manuscript?**

   Underlying cancer risk among patients with fatigue and other vague symptoms: a population-based cohort study in primary care.

   b) **Please include a link to or doi for the work**

   https://doi.org/10.3399/BJGP.2022.0371

   c) **Where was the work published?**

   British Journal of General Practice

   d) **Who published the work?** (e.g. OUP)

   Royal College of General Practitioners

   e) **When was the work published?**

   2023

   f) **List the manuscript's authors in the order they appear on the publication**

   White, B., Renzi, C., Barclay, M., & Lyratzopoulos, G.

   g) **Was the work peer reviewed?**

   Yes

   h) **Have you retained the copyright?**

   Yes

   i) **Was an earlier form of the manuscript uploaded to a preprint server?** (e.g. medRxiv). If 'Yes', please give a link or doi)

   No

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

☐

*I acknowledge permission of the publisher named under **1d** to include in this thesis portions of the publication named as included in **1c**.*

2. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)

BW, GL, and CR conceived and designed the study. BW managed and analysed the data, with statistical analyses and graphical presentation supervised/ developed by MB. CR and GL provided clinical input. All authors contributed to drafting and revising the article. Symptoms were defined using libraries of Read codes developed by Prof Willie Hamilton (WH) and Dr Sarah Price (SP) at Exeter University, with additional codes added by colleagues GL, CR, BW, MB, and Dr Meena Rafiq (MR) at UCL.

3. **In which chapter(s) of your thesis can this material be found?**

**5**

4. **e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

*Candidate*

*Date:*

05/06/2023

*Date*

*13/06/2023*

### 10.5.2 Potential cancer symptoms included in the study and sources of Read code lists used to define them.

Codes were amalgamated across sources, with minor additions by GL, CR, MB, MR, and BW. They were then deduplicated to ensure symptom definitions were mutually exclusive, and validated by clinical co-authors (GL, CR).

*Exeter University: Symptoms were defined using libraries of Read codes developed by WH and SP, and subsequently published in Moore et al 2021. UCL: Additional codes added following further empirical use by colleagues GL, CR, BW, MB, and MR.*

| Symptom | Source of Read code list |
|---|---|

| Alarm symptoms | |
|---|---|
| Abdominal (incl rectal) mass or intestinal obstruction | Renzi et al 2016 |
| Breast lump | Watson et al 2017, Exeter Uni., UCL |
| Breast skin changes | Din et al 2015 |
| Change in bowel habit | Watson et al 2017, Exeter Uni., UCL |
| Dysphagia | Watson et al 2017, Exeter Uni., UCL |
| Haematuria | Watson et al 2017, Exeter Uni., UCL |
| Haemoptysis | Watson et al 2017, Exeter Uni., UCL |
| Head or neck lump | Din et al 2015 |
| Hoarseness | Din et al 2015 |
| Jaundice | Watson et al 2017, Exeter Uni., UCL |
| Lump (excl. head & neck, breast, testicular, abdominal) | Din et al 2015 |
| Lymphadenopathy | Din et al 2015 |
| Nipple changes of concern (incl discharge or retraction) | Din et al 2015 |
| Post-menopausal bleeding | Watson et al 2017, Exeter Uni., UCL |
| Rectal bleeding | Watson et al 2017, Exeter Uni., UCL |
| Testicular enlargement/ lump | Din et al 2015 |
| **Vague symptoms** | |
| Abdominal bloating | Watson et al 2017, Exeter Uni., UCL |
| Abdominal pain | Watson et al 2017, Exeter Uni., UCL<br>Din et al 2015<br>Doran et al 2011 |
| Back pain | Doran et al 2011<br>Adapted by BW from Masefield et al 2021 |
| Chest pain | Din et al 2015 |
| Constipation | Renzi et al 2016 |
| Cough | Din et al 2015 |
| Diarrhoea | Renzi et al 2016 |
| Dyspnoea | Watson et al 2017, Exeter Uni., UCL |
| Fatigue | Exeter Uni., UCL |
| Headache | Adapted by BW from Masefield et al 2021 |
| Lower respiratory tract infection | Palin et al 2019 |
| Other musculoskeletal pain (excl back, pelvic, chest, abdominal, testicular) | Din et al 2015<br>Doran et al 2011<br>Adapted by BW from Masefield et al 2021 |
| Night sweats | Watson et al 2017, Exeter Uni., UCL |
| Pelvic pain | Din et al 2015<br>Zhou et al 2021 |

| | |
|---|---|
| Other upper gastro-intestinal (GI) symptoms (incl dyspepsia, nausea, vomiting, haematemesis, appetite loss) | Watson et al 2017, Exeter Uni., UCL<br>Zhou et al 2021<br>Bouras et al 2017<br>Din et al 2015 |
| Testicular pain | Din et al 2015 |
| Thromboembolic disease (incl pulmonary embolism) | Hawker et al 2014, available at Caliber PH674<br>Kuan et al 2019, available at Caliber PH338 v67<br>Kuan et al 2019, available at Caliber PH71 v142 |
| Upper respiratory tract infection | Palin et al 2019 |
| Urinary Tract Infections (incl cystitis, dysuria, urgency, painful urination, urine smell) | Zhou et al 2021 |
| Weight loss | Watson et al 2017, Exeter Uni., UCL |

### 10.5.3 Potential cancer symptoms excluded from the study due to unavailable Read code lists.

These potential cancer symptoms were included in NICE Guidelines or other sources (Rapid Diagnostic Centre Implementation Specification 2019, Chapman 2020, Chapman 2021), but could not be included in the study as Read code lists were not available.

| Symptom |
|---|
| **Alarm symptoms** |
| Vulval bleeding |
| Prostate feels malignant on examination |
| Anal mass or anal ulceration on examination |
| Vaginal mass |
| Vulval lump or ulceration |
| Appearance of cervix consistent with cervical cancer |
| Lip or oral cavity lump |
| Oral cavity red or red/ white patch erythroplakia or erythroleukoplakia |
| Alcohol induced lymph node pain |
| Skin lesion |
| Penile mass |
| Other penile symptoms affecting the foreskin or glans |
| **Vague symptoms** |
| Vaginal discharge |
| Oral cavity ulceration |
| Bleeding, bruising or petechiae |
| Pruritus |
| Splenomegaly |
| Hepatosplenomegaly |
| Irritable bowel syndrome |
| Finger clubbing |
| Loss of central neurological function |
| Chest signs consistent with lung cancer or pleural disease |
| Erectile dysfunction |
| Other testicular symptoms (excl lump, swelling, pain) |
| Fracture |
| Pallor |
| Infection (recurrent) |
| Fever |

### 10.5.4 Additional eligibility and validity criteria to define low haemoglobin

1. Patients with an eligible haemoglobin test

    a. First, we extracted all records from the Test file occurring 3 months before to 1 months after the patient's first fatigue presentation, where:
        i. EITHER the entity type was '173 - haemoglobin test'
        ii. OR the entity type was '288 – other' AND the record had a Read code for a haemoglobin test or anaemia (according to Read code lists developed by Din, 2015)

    b. The records were eligible for analysis if
        i. The record met the same criteria as other vague symptoms in the study i.e. the event date was:
            1. After the practice up to standard date (UTS)
            2. After the patient's current registration date to the practice (CRD)
            3. Before the last collection date for the practice (LCD)
            4. Before the patient's transfer out date from the practice (TOD) (if applicable)
            5. Before the patient's death date (if applicable)
            6. Before the patient was diagnosed with cancer (if applicable)
        ii. AND the test record had a Read term for haemoglobin or the Read term was missing (i.e. some records had a Read term for 'neutrophil count'- these were excluded)

    c. We then identified patients who had at least one eligible haemoglobin test conducted between 3 months before to 1 month after the first fatigue presentation.

2. Patients with valid haemoglobin test results

    a. We excluded haemoglobin results where the unit of measurement was invalid. Invalid units were: '%', 'mg/L', 'mmol/mol' , 'g/Kg' (see table below for possible measurement units used by previous publications).

    b. For haemoglobin results using a valid unit of measurement (g/dl or g/l), we excluded results if the values did not fall within the range of 'biologically plausible results' as previously defined by Caliber (https://www.caliberresearch.org/portal/show/haemoglobin_gprd):
        i. g/dl: >= 3 and < 25
        ii. g/l: >= 30 and < 250

    c. There were also haemoglobin results where the unit of measurement was not stated. These fell into two clusters, one where the unit of measurement was likely g/dl and one where the unit was likely g/l. We excluded these results, unless BOTH of the following criteria were met:
        i. The values fell within the range of 'biologically plausible results' for the two measurement units stated above
        ii. AND the test qualifier confirmed it was for a haemoglobin test

3. Patients with low haemoglobin

a. For valid haemoglobin results, we converted all results to the most frequent unit used, which was g/dl. Therefore, all g/l units were divided by 10.
b. We flagged records with a low haemoglobin result. This was defined based on values below the normal gender-specific range, provided by CPRD and available in published NICE Guidelines https://cks.nice.org.uk/topics/anaemia-iron-deficiency/diagnosis/investigations/):
    i. < 13 g/dl if male
    ii. <12 g/dl if female
c. For each patient, we identified the latest low haemoglobin value recorded within 3 months before the patient's first fatigue presentation, and the earliest low haemoglobin value recorded within 1 month afterwards.
d. We used this to flag patients who had **at least one low haemoglobin value** recorded between 3 months before to 1 month after the first fatigue presentation.

### 10.5.5 Combinations of different co-occurring vague symptoms and their cancer risk

a) Patients with combinations of 0, 1, or 2 or more different co-occurring vague symptoms, as a proportion of patients aged 30-99 years with fatigue and no alarm symptoms or anaemia

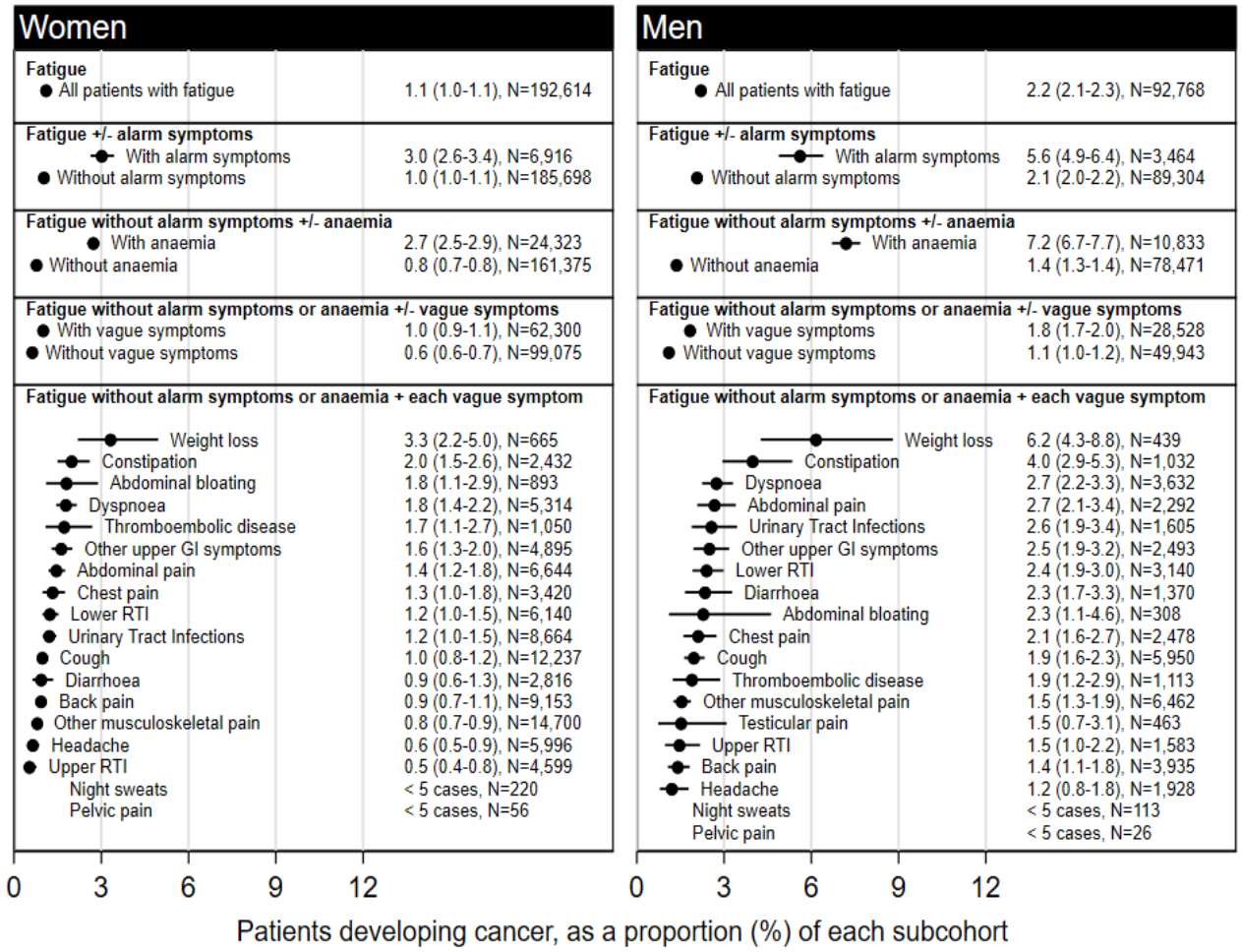
| | Male, binary 0/1 | | |
|---|---|---|---|
| | Women | Men | Total |
| Number of different additional vague symptoms | | | |
| 0 | | | |
| Freq. | 99,075 | 49,943 | 149,018 |
| % | 61.39 | 63.65 | 62.13 |
| 1 | | | |
| Freq. | 42,657 | 20,075 | 62,732 |
| % | 26.43 | 25.58 | 26.16 |
| 2+ | | | |
| Freq. | 19,643 | 8,453 | 28,096 |
| % | 12.17 | 10.77 | 11.71 |

b) Observed nine-month cancer risk (%) for patients aged 30-99 years with fatigue and no alarm symptoms or anaemia, by number of different co-occurring vague symptoms

| Number of different additional vague symptoms | Women | | | | Men | | | |
|---|---|---|---|---|---|---|---|---|
| | Total | Patients with cancer (n) | % | (95% CI) | Total | Patients with cancer (n) | % | (95% CI) |
| 0 | 99,075 | 613 | 0.62 | (0.57, 0.67) | 49,943 | 545 | 1.09 | (1.00, 1.19) |
| 1 | 42,657 | 355 | 0.83 | (0.75, 0.92) | 20,075 | 307 | 1.53 | (1.37, 1.71) |
| 2+ | 19,643 | 264 | 1.34 | (1.19, 1.51) | 8,453 | 212 | 2.51 | (2.20, 2.86) |

### 10.5.6 Graphs of observed cancer risk by each co-occurring symptom

Observed nine-month cancer risk (%) in patients aged 30-99 years with fatigue, by presence of each co-occurring symptom. Alarm symptoms do not include anaemia. Results for cohorts with under 5 cancer cases are suppressed to reduce statistical disclosure risk. Urinary Tract Infections also include cystitis, dysuria, urgency, painful urination, urine smell. Other upper GI (gastro-intestinal) symptoms include dyspepsia, nausea, vomiting, haematemesis, loss of appetite. RTI = Respiratory Tract Infections.



### 10.5.7 Incidence rate ratios of cancer for poisson models of co-occurring symptoms

Incidence rate ratios (IRR) for poisson models of nine-month cancer risk for patients presenting with fatigue, by presence of each co-occurring symptom, adjusting for age. Non-linear continuous year of age (30-99 years) was adjusted for using restricted cubic splines (results for age splines are not shown). Models were stratified by gender.

Separate models were run for:

1) patients with and without anaemia (restricted to patients with fatigue and no alarm symptom)
2) patients with and without any vague symptom (restricted to patients with fatigue and no alarm symptom or anaemia)
3) patients with and without each vague symptom (restricted to patients with fatigue and no alarm symptom or anaemia)

A) patients with and without anaemia (restricted to patients with fatigue and no alarm symptom)

| | Men | Women |
|---|---|---|
| Anaemia | | |
| IRR | 2.79 | 2.76 |
| 95% CI | (2.52 3.10) | (2.51 3.04) |

B) Patients with and without any vague symptom (restricted to patients with fatigue and no alarm symptom or anaemia)

| | Men | Women |
|---|---|---|
| Any vague symptom | | |
| IRR | 1.28 | 1.29 |
| 95% CI | (1.14 1.45) | (1.16 1.45) |

C) Patients with and without each vague symptom (restricted to patients with fatigue and no alarm symptom or anaemia)

Urinary Tract Infections also include cystitis, dysuria, urgency, painful urination, urine smell. Other upper GI (gastro-intestinal) symptoms include dyspepsia, nausea, vomiting, haematemesis, loss of appetite. RTI = Respiratory Tract Infections.

| | Men | Women |
|---|---|---|
| Abdominal pain | | |
| IRR | 1.70 | 1.84 |
| 95% CI | (1.31 2.20) | (1.48 2.28) |
| Abdominal bloating | | |
| IRR | 1.16 | 1.82 |
| 95% CI | (0.55 2.45) | (1.11 2.98) |
| Dyspnoea | | |
| IRR | 1.15 | 1.38 |
| 95% CI | (0.94 1.41) | (1.12 1.71) |
| Weight loss | | |
| IRR | 2.89 | 2.63 |

| | | |
|---|---|---|
| 95% CI | (2.00 4.19) | (1.73 3.99) |
| Constipation | | |
| IRR | 1.46 | 1.55 |
| 95% CI | (1.07 1.99) | (1.16 2.07) |
| Cough | | |
| IRR | 1.07 | 0.97 |
| 95% CI | (0.88 1.30) | (0.80 1.18) |
| Diarrhoea | | |
| IRR | 1.31 | 0.78 |
| 95% CI | (0.92 1.85) | (0.53 1.15) |
| Other upper GI symptoms | | |
| IRR | 1.49 | 1.54 |
| 95% CI | (1.15 1.94) | (1.22 1.95) |
| Urinary Tract Infections | | |
| IRR | 1.02 | 1.14 |
| 95% CI | (0.75 1.39) | (0.93 1.39) |
| Other musculoskeletal pain | | |
| IRR | 0.98 | 0.85 |
| 95% CI | (0.80 1.20) | (0.70 1.02) |
| Chest pain | | |
| IRR | 1.25 | 1.24 |
| 95% CI | (0.95 1.65) | (0.92 1.68) |
| Testicular pain | | |
| IRR | 1.19 | |
| 95% CI | (0.57 2.48) | |
| Headache | | |
| IRR | 1.08 | 1.03 |
| 95% CI | (0.71 1.63) | (0.74 1.42) |
| Back pain | | |
| IRR | 0.96 | 1.07 |
| 95% CI | (0.73 1.25) | (0.86 1.34) |
| Upper RTI | | |
| IRR | 1.05 | 0.70 |
| 95% CI | (0.70 1.59) | (0.47 1.05) |
| Lower RTI | | |
| IRR | 1.16 | 1.12 |
| 95% CI | (0.91 1.47) | (0.88 1.43) |
| Thromboembolic disease | | |
| IRR | 0.80 | 1.15 |
| 95% CI | (0.52 1.23) | (0.72 1.84) |

### 10.5.8 Table of modelled cancer risk with and without anaemia, by year of age

Modelled nine-month cancer risk (%) for patients presenting with fatigue without alarm symptoms, with and without anaemia, by year of age (30-99). Green = risk > 2%, orange = risk > 3%, red = risk > 6%. Includes observed nine-month cancer risk (%) for the general population in England in 2011, by five year age band. Available population estimates grouped all men/ women aged 85+. This table is available online at https://github.com/rmjlrwh/Fatigue.

Preview below:

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Male, binary 0/1 | Year of age | General popuation | With anaemia (%) | With anaemia (Upper 95% CI) | With anaemia (Lower 95% CI) | Without anaemia (%) | Without anaemia (Upper 95% CI) | Without anaemia (Lower 95% CI) |
| 2 | Women | 30 | 0.07 | 0.22 | 0.37 | 0.13 | 0.08 | 0.13 | 0.05 |
| 3 | Women | 31 | 0.07 | 0.23 | 0.37 | 0.14 | 0.08 | 0.13 | 0.05 |
| 4 | Women | 32 | 0.07 | 0.24 | 0.37 | 0.16 | 0.09 | 0.13 | 0.06 |
| 5 | Women | 33 | 0.07 | 0.26 | 0.37 | 0.18 | 0.09 | 0.14 | 0.07 |
| 6 | Women | 34 | 0.07 | 0.28 | 0.38 | 0.20 | 0.10 | 0.14 | 0.07 |
| 7 | Women | 35 | 0.11 | 0.29 | 0.39 | 0.22 | 0.11 | 0.14 | 0.08 |
| 8 | Women | 36 | 0.11 | 0.31 | 0.40 | 0.25 | 0.11 | 0.14 | 0.09 |
| 9 | Women | 37 | 0.11 | 0.34 | 0.42 | 0.27 | 0.12 | 0.15 | 0.10 |
| 10 | Women | 38 | 0.11 | 0.36 | 0.44 | 0.30 | 0.13 | 0.16 | 0.11 |
| 11 | Women | 39 | 0.11 | 0.39 | 0.47 | 0.32 | 0.14 | 0.17 | 0.12 |
| 12 | Women | 40 | 0.18 | 0.42 | 0.51 | 0.34 | 0.15 | 0.18 | 0.13 |
| 13 | Women | 41 | 0.18 | 0.45 | 0.55 | 0.37 | 0.16 | 0.20 | 0.13 |
| 14 | Women | 42 | 0.18 | 0.49 | 0.60 | 0.40 | 0.18 | 0.22 | 0.15 |
| 15 | Women | 43 | 0.18 | 0.54 | 0.66 | 0.44 | 0.19 | 0.24 | 0.16 |
| 16 | Women | 44 | 0.18 | 0.59 | 0.72 | 0.48 | 0.21 | 0.26 | 0.18 |
| 17 | Women | 45 | 0.29 | 0.65 | 0.78 | 0.54 | 0.24 | 0.28 | 0.20 |
| 18 | Women | 46 | 0.29 | 0.72 | 0.85 | 0.60 | 0.26 | 0.31 | 0.22 |
| 19 | Women | 47 | 0.29 | 0.79 | 0.93 | 0.67 | 0.29 | 0.33 | 0.25 |
| 20 | Women | 48 | 0.29 | 0.88 | 1.02 | 0.75 | 0.32 | 0.36 | 0.28 |
| 21 | Women | 49 | 0.29 | 0.97 | 1.12 | 0.84 | 0.35 | 0.40 | 0.31 |
| 22 | Women | 50 | 0.42 | 1.08 | 1.24 | 0.94 | 0.39 | 0.44 | 0.35 |

### 10.5.9 Table of modelled cancer risk with and without each vague symptom, by year of age

Modelled nine-month cancer risk (%) for patients presenting with fatigue without alarm symptoms or anaemia, with and without each vague symptom, by year of age (30-99). Green = risk > 2%, orange = risk > 3%, red = risk > 6%. Includes observed nine-month cancer risk (%) for the general population in England in 2011, by five year age band. Available population estimates grouped all men/ women aged 85+. Urinary Tract Infections also include cystitis, dysuria, urgency, painful urination, urine smell. Other upper GI (gastro-intestinal) symptoms include dyspepsia, nausea, vomiting, haematemesis, loss of appetite. RTI = Respiratory Tract Infections. This table is available online at https://github.com/rmjlrwh/Fatigue.

Preview below:

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Male, binary 0/1 | Year of age | General popuation | Without vague symptoms (%) | Without vague symptoms (Upper 95% CI) | Without vague symptoms (Lower 95% CI) | With vague symptoms (%) | With vague symptoms (Upper 95% CI) | With vague symptoms (Lower 95% CI) | Abdominal pain (%) | Abdominal pain (Upper 95% CI) | Abdominal pain (Lower 95% CI) |
| 2 | Women | 30 | 0.07 | 0.09 | 0.16 | 0.05 | 0.11 | 0.20 | 0.07 | 0.16 | 0.30 | 0.09 |
| 3 | Women | 31 | 0.07 | 0.09 | 0.15 | 0.06 | 0.12 | 0.20 | 0.07 | 0.17 | 0.29 | 0.10 |
| 4 | Women | 32 | 0.07 | 0.10 | 0.15 | 0.06 | 0.13 | 0.19 | 0.08 | 0.18 | 0.29 | 0.11 |
| 5 | Women | 33 | 0.07 | 0.11 | 0.15 | 0.07 | 0.13 | 0.19 | 0.09 | 0.19 | 0.30 | 0.13 |
| 6 | Women | 34 | 0.07 | 0.11 | 0.15 | 0.08 | 0.14 | 0.19 | 0.10 | 0.20 | 0.30 | 0.14 |
| 7 | Women | 35 | 0.11 | 0.12 | 0.16 | 0.09 | 0.15 | 0.20 | 0.11 | 0.22 | 0.30 | 0.15 |
| 8 | Women | 36 | 0.11 | 0.12 | 0.16 | 0.10 | 0.16 | 0.20 | 0.12 | 0.23 | 0.31 | 0.17 |
| 9 | Women | 37 | 0.11 | 0.13 | 0.16 | 0.10 | 0.17 | 0.21 | 0.13 | 0.24 | 0.33 | 0.18 |
| 10 | Women | 38 | 0.11 | 0.14 | 0.17 | 0.11 | 0.18 | 0.22 | 0.14 | 0.26 | 0.35 | 0.19 |
| 11 | Women | 39 | 0.11 | 0.15 | 0.18 | 0.12 | 0.19 | 0.23 | 0.15 | 0.27 | 0.37 | 0.20 |
| 12 | Women | 40 | 0.18 | 0.16 | 0.20 | 0.13 | 0.20 | 0.25 | 0.16 | 0.29 | 0.40 | 0.21 |
| 13 | Women | 41 | 0.18 | 0.17 | 0.21 | 0.14 | 0.21 | 0.27 | 0.17 | 0.31 | 0.43 | 0.23 |
| 14 | Women | 42 | 0.18 | 0.18 | 0.23 | 0.15 | 0.23 | 0.29 | 0.18 | 0.33 | 0.46 | 0.24 |

### 10.5.10 Frequency of the three most common cancer sites, by co-occurring symptom

Proportion of patients diagnosed with cancer whose first cancer was one of the three most commonly diagnosed cancers in each co-occurring vague symptom group. Includes patients with fatigue aged 30-99 years who had each co-occurring symptom 3 months before to 1 month after the first fatigue presentation (restricted to patients with fatigue and no alarm symptom or anaemia). N/a = Analysis excluded symptom combinations with no cancer cases.

| Pairwise combinations of fatigue with each vague symptom | Total number of patients diagnosed with cancer within 9 months (N) | Proportion of patients with cancer whose first cancer diagnosis was one of the three most common cancer sites (%) |
|---|---|---|
| Men | | |
| Night sweats | < 5 | 100.0 |
| Abdominal bloating | 7 | 85.7 |
| Dyspnoea | 99 | 68.7 |
| Other musculoskeletal pain | 99 | 65.7 |
| Upper RTI | 23 | 65.2 |
| Lower RTI | 75 | 64.0 |
| Weight loss | 27 | 63.0 |
| Thromboembolic disease | 21 | 61.9 |
| Headache | 23 | 60.9 |
| Cough | 116 | 60.3 |
| Constipation | 41 | 58.5 |
| Back pain | 55 | 58.2 |
| Other upper GI symptoms | 62 | 58.1 |
| Chest pain | 52 | 55.8 |
| Urinary Tract Infections | 41 | 51.2 |
| Diarrhoea | 32 | 50.0 |
| Abdominal pain | 61 | 45.9 |
| Testicular pain | 7 | 42.9 |
| Pelvic pain | N/a | N/a |
| Women | | |
| Night sweats | < 5 | 100.0 |
| Weight loss | 22 | 63.6 |
| Cough | 118 | 59.3 |
| Dyspnoea | 94 | 58.5 |
| Upper RTI | 24 | 58.3 |
| Diarrhoea | 26 | 57.7 |
| Lower RTI | 75 | 57.3 |
| Abdominal bloating | 16 | 56.3 |
| Constipation | 48 | 56.3 |
| Back pain | 84 | 56.0 |
| Thromboembolic disease | 18 | 55.6 |
| Headache | 38 | 52.6 |
| Chest pain | 45 | 46.7 |
| Other musculoskeletal pain | 116 | 45.7 |
| Urinary Tract Infections | 104 | 42.3 |
| Abdominal pain | 96 | 41.7 |
| Other upper GI symptoms | 79 | 38.0 |
| Pelvic pain | N/a | N/a |

## 10.6 Chapter 6 appendices

### 10.6.1 UCL Research paper declaration form

# UCL Research Paper Declaration Form
## referencing the doctoral candidate's own published work(s)

*Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:*

- *have been uploaded to a preprint server*
- *are in submission to a peer-reviewed publication*
- *have been published in a peer-reviewed publication, e.g. journal, textbook.*

*This form should be completed as many times as necessary. For instance, if you have seven thesis chapters, two of which containing material that has already been published, you would complete this form twice.*

1. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)

   a) **What is the current title of the manuscript?**

   Incident disease among patients presenting in primary care with fatigue: a population-based cohort study

   b) **Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv; if 'Yes', please give a link or doi)

   No

   c) **Where is the work intended to be published?** (e.g. journal names)

   TBC

   d) **List the manuscript's authors in the intended authorship order**

   Becky White, Nadine Zakkak, Cristina Renzi, Meena Rafiq, Arturo Gonzalez-Izquierdo, Spiros Denaxas, Brian D Nicholson, Georgios Lyratzopoulos, Matthew Barclay

   e) **Stage of publication** (e.g. in submission)

   In preparation

2. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)

   BW, MB, and GL conceived and designed the study. BW managed and analysed the data, under the supervision of MB, who also provided statistical expertise and designed the DAG diagrams in the discussion. MB and NZ shared analytical code used to manage and analyse the data, and NZ quality assured the final code and outputs. BDN, MR, CR, and GL provided clinical input, and AGI and SD developed disease phenotypes and advised on the presentation and discussion of results. All authors contributed to drafting and revising the article.

3. **In which chapter(s) of your thesis can this material be found?**

   **6**

4. **e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

   *Candidate*

   *Date:*

   05/06/2023

   *Date*

   *13/06/2023*

   Click or tap here to enter text.

### 10.6.2 Fatigue phenotype

This table is available online at https://github.com/rmjlrwh/FatigueRiskMap

| medcode | readcode | readcode_desc |
|---|---|---|
| 1042 | R007400 | [D]Postviral (asthenic) syndrome |
| 1147 | R007500 | [D]Tiredness |
| 1371 | R007300 | [D]Lethargy |
| 1404 | 1682.00 | Fatigue |
| 1582 | E205.11 | Nervous exhaustion |
| 1688 | R007100 | [D]Fatigue |
| 1900 | R2y3.00 | "[D]Debility, unspecified" |
| 2855 | 1B32.00 | Weakness present |
| 3361 | E205.00 | Neurasthenia - nervous debility |
| 4546 | F286.00 | Chronic fatigue syndrome |
| 5049 | R007200 | [D]Asthenia NOS |
| 5583 | 168..12 | Lethargy - symptom |
| 5658 | R007000 | [D]Malaise |
| 5751 | 1683.00 | Tired all the time |
| 5794 | 168..00 | Tiredness symptom |
| 5814 | R007z11 | [D]Lassitude |
| 6029 | 1B3..12 | Weakness symptoms |
| 6190 | F286.12 | Postviral fatigue syndrome |
| 6242 | 168..11 | Fatigue - symptom |
| 7235 | E205.12 | Tired all the time |
| 7529 | F286.11 | CFS - Chronic fatigue syndrome |
| 9127 | F286.14 | Post-viral fatigue syndrome |
| 9220 | 1688.00 | Exhaustion |
| 9435 | 2254.00 | O/E - apathetic |
| 9656 | Eu46011 | [X]Fatigue syndrome |
| 9823 | 1684.11 | C/O - debility - malaise |
| 9889 | R007211 | [D]General weakness |
| 12411 | R007411 | [D]Post viral debility |
| 15516 | 1683.11 | C/O - 'tired all the time' |
| 16479 | 2832.12 | O/E - weakness |
| 16561 | Eu46000 | [X]Neurasthenia |
| 17736 | 1684.00 | Malaise/lethargy |
| 23932 | R007z00 | [D]Malaise and fatigue NOS |
| 24382 | R204.00 | [D]Senile exhaustion |
| 27877 | F286.13 | PVFS - Postviral fatigue syn |
| 29292 | 168Z.00 | Tiredness symptom NOS |
| 44215 | R007.00 | [D]Malaise and fatigue |
| 97284 | F286100 | Moderate chronic fatigue syndrome |
| 98512 | F286000 | Mild chronic fatigue syndrome |

| | | |
|---|---|---|
| 98734 | F286200 | Severe chronic fatigue syndrome |
| 99807 | 8HkW.00 | Referral to chronic fatigue syndrome specialist team |
| 97140 | 8Q1..00 | Activity management for chronic fatigue syndrome |
| 1816 | 168..13 | Malaise - symptom |

### 10.6.3 List of included conditions and published phenotype sources

This table is available online at https://github.com/rmjlrwh/FatigueRiskMap

| Disease number | Disease name | Disease group | CPRD Read codes used | HES APC ICD 10 codes used | CR ICD 10 codes used | HES APC OPCS codes used | Primary code list source was Kuan et al | 2nd code list source | 3rd code list source | 4th code list source | Period used to capture previous diagnoses |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Benign neoplasm and polyp of uterus | Benign Neoplasm/CIN | 1 | 1 | | | y | | | | ever |
| 2 | Benign neoplasm of colon, rectum, anus and anal canal | Benign Neoplasm/CIN | 1 | 1 | | | y | | | | ever |
| 3 | Benign neoplasm of ovary | Benign Neoplasm/CIN | 1 | 1 | | | y | | | | ever |
| 4 | Benign neoplasm of stomach and duodenum | Benign Neoplasm/CIN | 1 | 1 | | | y | | | | ever |
| 5 | Haemangioma, any site | Benign Neoplasm/CIN | 1 | 1 | | | y | | | | ever |
| 6 | Leiomyoma of uterus | Benign Neoplasm/CIN | 1 | 1 | | | y | | | | ever |
| 7 | Breast cancer | Cancers | | | 1 | | | Adapted from NHS England's National Cancer Waiting Times Monitoring dataset | | | ever |
| 8 | Gynaecological cancer | Cancers | | | 1 | | | Adapted from NHS England's National Cancer Waiting Times Monitoring dataset | | | ever |
| 9 | Haematological cancer | Cancers | | | 1 | | | Adapted from NHS England's National Cancer Waiting Times Monitoring dataset | | | ever |
| 10 | Lower GI cancer | Cancers | | | 1 | | | Adapted from NHS England's National Cancer | | | ever |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Waiting Times Monitoring dataset | | | |
| 11 | Lung cancer | Cancers | | | 1 | | | Adapted from NHS England's National Cancer Waiting Times Monitoring dataset | | | ever |
| 12 | Other cancer | Cancers | | | 1 | | | Adapted from NHS England's National Cancer Waiting Times Monitoring dataset | | | ever |
| 13 | Prostate cancer | Cancers | | | 1 | | | Adapted from NHS England's National Cancer Waiting Times Monitoring dataset | | | ever |
| 14 | Upper GI cancer | Cancers | | | 1 | | | Adapted from NHS England's National Cancer Waiting Times Monitoring dataset | | | ever |
| 15 | Urological cancer | Cancers | | | 1 | | | Adapted from NHS England's National Cancer Waiting Times Monitoring dataset | | | ever |
| 16 | Abdominal aortic aneurysm | Diseases of the Cardiovascular system | | 1 | | | y | | | | ever |
| 17 | Atrial fibrillation | Diseases of the Cardiovascular system | | 1 | | | y | | | | ever |
| 18 | Atrioventricular block, complete | Diseases of the Cardiovascular system | 1 | 1 | | | y | | | | ever |
| 19 | Atrioventricular block, first degree | Diseases of the Cardiovascular system | 1 | 1 | | | y | | | | ever |
| 20 | Atrioventricular block, second degree | Diseases of the Cardiovascular system | 1 | 1 | | | y | | | | ever |
| 21 | Bifascicular block | Diseases of the Cardiovascular system | 1 | 1 | | | y | | | | ever |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 22 | Coronary heart disease not otherwise specified | Diseases of the Cardiovascular system | | 1 | | | y | | | | ever |
| 23 | Dilated cardiomyopathy | Diseases of the Cardiovascular system | 1 | 1 | | | y | | | | ever |
| 24 | Heart failure | Diseases of the Cardiovascular system | 1 | 1 | | | y | | | | ever |
| 25 | Hypertension | Diseases of the Cardiovascular system | 1 | 1 | | | y | | | | ever |
| 26 | Hypertrophic Cardiomyopathy | Diseases of the Cardiovascular system | 1 | 1 | | | y | | | | ever |
| 27 | Intracerebral haemorrhage | Diseases of the Cardiovascular system | 1 | 1 | | | y | | | | ever |
| 28 | Ischaemic stroke | Diseases of the Cardiovascular system | 1 | 1 | | | y | | | | ever |
| 29 | Left bundle branch block | Diseases of the Cardiovascular system | 1 | 1 | | | y | | | | ever |
| 30 | Multiple valve dz | Diseases of the Cardiovascular system | 1 | 1 | | | y | | | | ever |
| 31 | Myocardial infarction | Diseases of the Cardiovascular system | | 1 | | | y | | | | ever |
| 32 | Nonrheumatic aortic valve disorders | Diseases of the Cardiovascular system | 1 | 1 | | | y | | | | ever |
| 33 | Nonrheumatic mitral valve disorders | Diseases of the Cardiovascular system | 1 | 1 | | | y | | | | ever |
| 34 | Other Cardiomyopathy | Diseases of the Cardiovascular system | 1 | 1 | | | y | | | | ever |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 35 | Pericardial effusion (noninflammatory) | Diseases of the Cardiovascular system | 1 | 1 | | | y | | | | ever |
| 36 | Peripheral arterial disease | Diseases of the Cardiovascular system | | 1 | | 1 | y | | | | ever |
| 37 | Primary pulmonary hypertension | Diseases of the Cardiovascular system | 1 | 1 | | | y | | | | ever |
| 38 | Pulmonary embolism | Diseases of the Cardiovascular system | 1 | 1 | | | y | | | | ever |
| 39 | Rheumatic valve dz | Diseases of the Circulatory System | 1 | 1 | | | y | | | | ever |
| 40 | Right bundle branch block | Diseases of the Circulatory System | 1 | 1 | | | y | | | | ever |
| 41 | Sick sinus syndrome | Diseases of the Circulatory System | 1 | 1 | | | y | | | | ever |
| 42 | Stable angina | Diseases of the Circulatory System | | 1 | | | y | | | | ever |
| 43 | Stroke NOS | Diseases of the Circulatory System | 1 | 1 | | | y | | | | ever |
| 44 | Subarachnoid haemorrhage | Diseases of the Circulatory System | 1 | 1 | | | y | | | | ever |
| 45 | Subdural haematoma - nontraumatic | Diseases of the Circulatory System | 1 | 1 | | | y | | | | ever |
| 46 | Supraventricular tachycardia | Diseases of the Circulatory System | 1 | 1 | | | y | | | | ever |
| 47 | Transient ischaemic attack | Diseases of the Circulatory System | 1 | 1 | | | y | | | | ever |
| 48 | Trifascicular block | Diseases of the Circulatory System | 1 | 1 | | | y | | | | ever |
| 49 | Unstable Angina | Diseases of the Circulatory System | | 1 | | | y | | | | ever |
| 50 | Venous thromboembolic disease (Excl PE) | Diseases of the Circulatory System | 1 | 1 | | | y | | | | ever |
| 51 | Ventricular tachycardia | Diseases of the Circulatory System | 1 | 1 | | | y | | | | ever |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 52 | Abdominal Hernia | Diseases of the Digestive System | 1 | 1 | | 1 | y | | | | ever |
| 53 | Alcoholic liver disease | Diseases of the Digestive System | 1 | 1 | | | y | | | | ever |
| 54 | Anal fissure | Diseases of the Digestive System | 1 | 1 | | 1 | y | | | | ever |
| 55 | Angiodysplasia of colon | Diseases of the Digestive System | 1 | 1 | | | y | | | | ever |
| 56 | Anorectal fistula | Diseases of the Digestive System | 1 | 1 | | 1 | y | | | | ever |
| 57 | Anorectal prolapse | Diseases of the Digestive System | 1 | 1 | | 1 | y | | | | ever |
| 58 | Appendicitis | Diseases of the Digestive System | 1 | 1 | | 1 | y | | | | ever |
| 59 | Autoimmune liver disease | Diseases of the Digestive System | 1 | 1 | | | y | | | | ever |
| 60 | Barrett's oesophagus | Diseases of the Digestive System | 1 | 1 | | | y | | | | ever |
| 61 | Cholangitis | Diseases of the Digestive System | 1 | 1 | | | y | | | | ever |
| 62 | Cholecystitis | Diseases of the Digestive System | 1 | 1 | | | y | | | | ever |
| 63 | Cholelithiasis | Diseases of the Digestive System | 1 | 1 | | | y | | | | ever |
| 64 | Coeliac disease | Diseases of the Digestive System | 1 | 1 | | | y | | | | ever |
| 65 | Diaphragmatic hernia | Diseases of the Digestive System | 1 | 1 | | 1 | y | | | | ever |
| 66 | Diverticular disease of intestine (acute and chronic) | Diseases of the Digestive System | 1 | 1 | | 1 | y | | | | ever |
| 67 | Fatty Liver | Diseases of the Digestive System | 1 | 1 | | | y | | | | ever |
| 68 | Gastritis and duodenitis | Diseases of the Digestive System | 1 | 1 | | | y | | | | ever |
| 69 | Gastro-oesophageal reflux disease | Diseases of the Digestive System | 1 | 1 | | 1 | y | | | | ever |
| 70 | Hepatic failure | Diseases of the Digestive System | 1 | 1 | | | y | | | | ever |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 71 | Inflammatory Bowel Disease | Diseases of the Digestive System | 1 | 1 | | | y | | | | ever |
| 73 | Irritable bowel syndrome | Diseases of the Digestive System | 1 | 1 | | | y | | | | ever |
| 74 | Liver fibrosis, sclerosis and cirrhosis | Diseases of the Digestive System | 1 | 1 | | | y | | | | ever |
| 75 | Oesophagitis and oesophageal ulcer | Diseases of the Digestive System | 1 | 1 | | | y | | | | ever |
| 76 | Pancreatitis | Diseases of the Digestive System | 1 | 1 | | | y | | | | ever |
| 77 | Peptic ulcer disease | Diseases of the Digestive System | 1 | 1 | | 1 | y | | | | ever |
| 78 | Peritonitis | Diseases of the Digestive System | 1 | 1 | | 1 | y | | | | ever |
| 79 | Portal hypertension | Diseases of the Digestive System | 1 | 1 | | 1 | y | | | | ever |
| 81 | Volvulus | Diseases of the Digestive System | 1 | 1 | | 1 | y | | | | ever |
| 82 | Hearing loss | Diseases of the Ear | 1 | 1 | | | y | | | | ever |
| 83 | Meniere disease | Diseases of the Ear | 1 | 1 | | | y | | | | ever |
| 84 | Tinnitus | Diseases of the Ear | 1 | 1 | | | y | | | | ever |
| 85 | Diabetes | Diseases of the Endocrine System | 1 | 1 | | | y | | | | ever |
| 86 | Hyperparathyroidism | Diseases of the Endocrine System | 1 | 1 | | | y | | | | ever |
| 87 | Hypo or hyperthyroidism | Diseases of the Endocrine System | 1 | 1 | | | y | | | | ever |
| 88 | Polycystic ovarian syndrome | Diseases of the Endocrine System | 1 | 1 | | | y | | | | ever |
| 89 | Syndrome of inappropriate secretion of antidiuretic hormone | Diseases of the Endocrine System | | 1 | | | y | | | | ever |
| 90 | Anterior and Intermediate Uveitis | Diseases of the Eye | 1 | 1 | | | y | | | | ever |
| 91 | Cataract | Diseases of the Eye | 1 | 1 | | 1 | y | | | | ever |
| 92 | Diabetic ophthalmic complications | Diseases of the Eye | 1 | 1 | | | y | | | | ever |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 93 | Glaucoma | Diseases of the Eye | 1 | 1 | | 1 | y | | | | ever |
| 94 | Keratitis | Diseases of the Eye | 1 | 1 | | | y | | | | ever |
| 95 | Macular degeneration | Diseases of the Eye | 1 | 1 | | | y | | | | ever |
| 96 | Posterior Uveitis | Diseases of the Eye | 1 | 1 | | | y | | | | ever |
| 97 | Ptosis of eyelid | Diseases of the Eye | 1 | 1 | | 1 | y | | | | ever |
| 98 | Retinal detachments and breaks | Diseases of the Eye | 1 | 1 | | 1 | y | | | | ever |
| 99 | Retinal vascular occlusions | Diseases of the Eye | 1 | 1 | | | y | | | | ever |
| 100 | Scleritis and episcleritis | Diseases of the Eye | 1 | 1 | | | y | | | | ever |
| 101 | Visual impairment and blindness | Diseases of the Eye | 1 | 1 | | | y | | | | ever |
| 102 | Acute Kidney Injury | Diseases of the Genitourinary system | | 1 | | | y | | | | ever |
| 103 | Chronic kidney disease | Diseases of the Genitourinary system | 1 | 1 | | 1 | | Kontopantelis 2015 | Knight et al https://phenotypes.healthdatagateway.org./phenotypes/PH950/version/2128/detail | | ever |
| 104 | Dysmenorrhoea | Diseases of the Genitourinary system | 1 | 1 | | | y | | | | ever |
| 105 | Endometrial hyperplasia and hypertrophy | Diseases of the Genitourinary system | 1 | 1 | | | y | | | | ever |
| 106 | Endometriosis | Diseases of the Genitourinary system | 1 | 1 | | | y | | | | ever |
| 107 | Erectile dysfunction | Diseases of the Genitourinary system | 1 | 1 | | | y | | | | ever |
| 108 | Female genital prolapse | Diseases of the Genitourinary system | 1 | 1 | | | y | | | | ever |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 109 | Female infertility | Diseases of the Genitourinary system | 1 | 1 | | | y | | | | ever |
| 110 | Glomerulonephritis | Diseases of the Genitourinary system | 1 | 1 | | | y | | | | ever |
| 111 | Hydrocoele (incl infected) | Diseases of the Genitourinary system | 1 | 1 | | | y | | | | ever |
| 112 | Hyperplasia of prostate | Diseases of the Genitourinary system | 1 | 1 | | | y | | | | ever |
| 113 | Male infertility | Diseases of the Genitourinary system | 1 | 1 | | | y | | | | ever |
| 114 | Menorrhagia and polymenorrhoea | Diseases of the Genitourinary system | 1 | 1 | | | y | | | | ever |
| 115 | Neuromuscular dysfunction of bladder | Diseases of the Genitourinary system | 1 | 1 | | | y | | | | ever |
| 116 | Non-acute cystitis | Diseases of the Genitourinary system | 1 | 1 | | | y | | | | ever |
| 117 | Obstructive and reflux uropathy | Diseases of the Genitourinary system | 1 | 1 | | | y | | | | ever |
| 118 | Tubulo-interstitial nephritis | Diseases of the Genitourinary system | 1 | 1 | | | y | | | | ever |
| 119 | Urinary Incontinence | Diseases of the Genitourinary system | 1 | 1 | | | y | | | | ever |
| 120 | Urolithiasis | Diseases of the Genitourinary system | 1 | 1 | | 1 | y | | | | ever |
| 121 | Allergic and chronic rhinitis | Diseases of the Respiratory System | 1 | 1 | | | y | | | | ever |
| 122 | Asbestosis | Diseases of the Respiratory System | 1 | 1 | | | y | | | | ever |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 123 | Aspiration pneumonitis | Diseases of the Respiratory System | 1 | 1 | | | y | | | | ever |
| 124 | Asthma | Diseases of the Respiratory System | 1 | 1 | | | y | | | | ever |
| 125 | Bronchiectasis | Diseases of the Respiratory System | 1 | 1 | | | y | | | | ever |
| 126 | COPD | Diseases of the Respiratory System | 1 | 1 | | | y | | | | ever |
| 127 | Chronic sinusitis | Diseases of the Respiratory System | 1 | 1 | | | y | | | | ever |
| 128 | Hypertrophy of nasal turbinates | Diseases of the Respiratory System | 1 | 1 | | | y | | | | ever |
| 129 | Nasal polyp | Diseases of the Respiratory System | 1 | 1 | | | y | | | | ever |
| 130 | Other interstitial pulmonary diseases with fibrosis | Diseases of the Respiratory System | 1 | 1 | | | y | | | | ever |
| 131 | Pleural effusion | Diseases of the Respiratory System | 1 | 1 | | | y | | | | ever |
| 132 | Pleural plaque | Diseases of the Respiratory System | 1 | 1 | | | y | | | | ever |
| 133 | Pneumonitis | Diseases of the Respiratory System | | 1 | | | | Ramirez 2018 https://phekb.org/phenotype/pneumonia-vumc-emerge-v51 | | | ever |
| 134 | Pneumothorax | Diseases of the Respiratory System | 1 | 1 | | | y | | | | ever |
| 135 | Respiratory failure | Diseases of the Respiratory System | 1 | 1 | | | y | | | | ever |
| 136 | Sleep apnoea | Diseases of the Respiratory System | 1 | 1 | | | y | | | | ever |
| 137 | Agranulocytosis | Haematological/Immunological conditions | 1 | 1 | | | y | | | | ever |
| 138 | Aplastic anaemias | Haematological/Immunological conditions | 1 | 1 | | | y | | | | ever |
| 139 | Folate deficiency anaemia | Haematological/Immunological conditions | 1 | 1 | | | y | | | | ever |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 140 | Immunodeficiencies | Haematological/Immunological conditions | 1 | 1 | | | y | | | | ever |
| 141 | Iron deficiency anaemia | Haematological/Immunological conditions | 1 | 1 | | | y | | | | ever |
| 142 | Other anaemias | Haematological/Immunological conditions | 1 | 1 | | | y | | | | ever |
| 143 | Other haemolytic anaemias | Haematological/Immunological conditions | 1 | 1 | | | y | | | | ever |
| 144 | Raynaud's syndrome | Haematological/Immunological conditions | 1 | 1 | | | y | | | | ever |
| 145 | Rheumatoid Arthritis | Haematological/Immunological conditions | 1 | 1 | | | y | | | | ever |
| 146 | Sarcoidosis | Haematological/Immunological conditions | 1 | 1 | | | y | | | | ever |
| 147 | Sickle-cell anaemia | Haematological/Immunological conditions | 1 | 1 | | | y | | | | ever |
| 148 | Sjogren's disease | Haematological/Immunological conditions | 1 | 1 | | | y | | | | ever |
| 149 | Thalassaemia | Haematological/Immunological conditions | 1 | 1 | | | y | | | | ever |
| 150 | Vitamin B12 deficiency anaemia | Haematological/Immunological conditions | 1 | 1 | | | y | | | | ever |
| 151 | Bacterial Diseases (excl TB) | Infectious Diseases | | 1 | | | y | | | | 2yr |
| 152 | Chronic viral hepatitis | Infectious Diseases | 1 | 1 | | | y | | | | ever |
| 153 | Ear and Upper Respiratory Tract Infections | Infectious Diseases | 1 | 1 | | | y | | | | 2yr |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 154 | Encephalitis | Infectious Diseases | | 1 | | | y | | | | 2yr |
| 155 | Eye infections | Infectious Diseases | 1 | 1 | | | y | | | | 2yr |
| 156 | Female pelvic inflammatory disease | Infectious Diseases | | 1 | | | y | | | | 2yr |
| 157 | HIV | Infectious Diseases | 1 | 1 | | | y | | | | ever |
| 158 | Infection of anal and rectal regions | Infectious Diseases | | 1 | | | y | | | | 2yr |
| 159 | Infection of bones and joints | Infectious Diseases | | 1 | | | y | | | | 2yr |
| 160 | Infection of liver | Infectious Diseases | | 1 | | | y | | | | 2yr |
| 161 | Infection of male genital system | Infectious Diseases | | 1 | | | y | | | | 2yr |
| 162 | Infection of other or unspecified genitourinary system | Infectious Diseases | | 1 | | | y | | | | 2yr |
| 163 | Infection of skin and subcutaneous tissues | Infectious Diseases | | 1 | | | y | | | | 2yr |
| 164 | Infections of Other or unspecified organs | Infectious Diseases | | 1 | | | y | | | | 2yr |
| 165 | Infections of the Heart | Infectious Diseases | | 1 | | | y | | | | 2yr |
| 166 | Infections of the digestive system | Infectious Diseases | | 1 | | | y | | | | 2yr |
| 167 | Lower Respiratory Tract Infections | Infectious Diseases | 1 | 1 | | | y | | | | 2yr |
| 168 | Lyme disease | Infectious Diseases | 1 | 1 | | | | Cairns 2019 bmjopen.bmj.com/content/9/7/e025916 | Emma Whitfield 2022 github.com/ekw26/Atlas-phenotypes | | 2yr |
| 169 | Meningitis | Infectious Diseases | | 1 | | | y | | | | 2yr |
| 170 | Mycoses | Infectious Diseases | | 1 | | | y | | | | 2yr |
| 171 | Other nervous system infections | Infectious Diseases | | 1 | | | y | | | | 2yr |
| 172 | Other or unspecified infectious organisms | Infectious Diseases | | 1 | | | y | | | | 2yr |
| 173 | Parasitic infections | Infectious Diseases | | 1 | | | y | | | | 2yr |
| 174 | Rheumatic fever | Infectious Diseases | 1 | 1 | | | y | | | | ever |
| 175 | Septicaemia | Infectious Diseases | | 1 | | | y | | | | 2yr |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 176 | Tuberculosis | Infectious Diseases | 1 | 1 | | | y | | | | ever |
| 177 | Urinary Tract Infections | Infectious Diseases | 1 | 1 | | | y | | | | 2yr |
| 178 | Viral diseases (excl chronic hepatitis/HIV) | Infectious Diseases | | 1 | | | y | | | | 2yr |
| 179 | Alcohol Problems | Mental Health Disorders | 1 | 1 | | | y | | | | ever |
| 180 | Anorexia and bulimia nervosa | Mental Health Disorders | 1 | 1 | | | y | | | | ever |
| 181 | Anxiety disorders | Mental Health Disorders | 1 | 1 | | | y | | | | ever |
| 182 | Autism and Asperger's syndrome | Mental Health Disorders | 1 | 1 | | | y | | | | ever |
| 183 | Bipolar affective disorder and mania | Mental Health Disorders | 1 | 1 | | | y | | | | ever |
| 184 | Delirium, not induced by alcohol and other psychoactive substances | Mental Health Disorders | 1 | 1 | | | y | | | | ever |
| 185 | Depression | Mental Health Disorders | 1 | 1 | | | y | | | | ever |
| 186 | Hyperkinetic disorders | Mental Health Disorders | 1 | 1 | | | y | | | | ever |
| 187 | Insomnia & sleep disturbances | Mental Health Disorders | 1 | 1 | | | | Moore 2021: https://www.nature.com/articles/s41416-021-01541-4 | Hoile 2019 https://clinicalcodes.rss.mhs.man.ac.uk/medcodes/article/78/codelist/res78-insomnia/ | | ever |
| 188 | Obsessive-compulsive disorder | Mental Health Disorders | 1 | 1 | | | y | | | | ever |
| 189 | Other psychoactive substance misuse | Mental Health Disorders | 1 | 1 | | | y | | | | ever |
| 190 | Personality disorders | Mental Health Disorders | 1 | 1 | | | y | | | | ever |
| 191 | Schizophrenia, schizotypal and delusional disorders | Mental Health Disorders | 1 | 1 | | | y | | | | ever |
| 192 | Ankylosing spondylitis | Musculoskeletal conditions | 1 | 1 | | | y | | | | ever |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 193 | Carpal tunnel syndrome | Musculoskeletal conditions | 1 | 1 | | 1 | y | | | | ever |
| 194 | Collapsed vertebra | Musculoskeletal conditions | 1 | 1 | | 1 | y | | | | ever |
| 195 | Connective & soft tissue disorders | Musculoskeletal conditions | 1 | 1 | | | y | | | | ever |
| 196 | Enteropathic arthropathy | Musculoskeletal conditions | 1 | 1 | | | y | | | | ever |
| 197 | Fibromatoses | Musculoskeletal conditions | 1 | 1 | | 1 | y | | | | ever |
| 198 | Fracture of hip | Musculoskeletal conditions | 1 | 1 | | 1 | y | | | | ever |
| 199 | Fracture of wrist | Musculoskeletal conditions | 1 | 1 | | | y | | | | ever |
| 200 | Giant Cell arteritis | Musculoskeletal conditions | 1 | 1 | | | y | | | | ever |
| 201 | Gout | Musculoskeletal conditions | 1 | 1 | | | y | | | | ever |
| 202 | Intervertebral disc disorders | Musculoskeletal conditions | 1 | 1 | | 1 | y | | | | ever |
| 203 | Lupus erythematosus (local and systemic) | Musculoskeletal conditions | 1 | 1 | | | y | | | | ever |
| 204 | Osteoarthritis (excl spine) | Musculoskeletal conditions | 1 | 1 | | | y | | | | ever |
| 205 | Osteoporosis | Musculoskeletal conditions | 1 | 1 | | | y | | | | ever |
| 206 | Polymyalgia Rheumatica | Musculoskeletal conditions | 1 | 1 | | | y | | | | ever |
| 207 | Postinfective and reactive arthropathies | Musculoskeletal conditions | 1 | 1 | | | y | | | | ever |
| 208 | Psoriatic arthropathy | Musculoskeletal conditions | 1 | 1 | | | y | | | | ever |
| 209 | Scoliosis | Musculoskeletal conditions | 1 | 1 | | | y | | | | ever |
| 210 | Spinal stenosis | Musculoskeletal conditions | 1 | 1 | | | y | | | | ever |
| 211 | Spondylolisthesis | Musculoskeletal conditions | 1 | 1 | | | y | | | | ever |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 212 | Spondylosis | Musculoskeletal conditions | 1 | 1 | | | y | | | | ever |
| 213 | Systemic sclerosis | Musculoskeletal conditions | 1 | 1 | | | y | | | | ever |
| 214 | Bell's palsy | Neurological conditions | 1 | 1 | | | y | | | | ever |
| 215 | Dementia | Neurological conditions | 1 | 1 | | | y | | | | ever |
| 216 | Disorders of autonomic nervous system | Neurological conditions | 1 | 1 | | | y | | | | ever |
| 217 | Epilepsy | Neurological conditions | 1 | 1 | | | y | | | | ever |
| 218 | Essential tremor | Neurological conditions | 1 | 1 | | | y | | | | ever |
| 219 | Intracranial hypertension | Neurological conditions | 1 | 1 | | | y | | | | ever |
| 220 | Migraine | Neurological conditions | 1 | 1 | | | y | | | | ever |
| 221 | Motor neuron disease | Neurological conditions | 1 | 1 | | | y | | | | ever |
| 222 | Multiple sclerosis | Neurological conditions | 1 | 1 | | | y | | | | ever |
| 223 | Myasthenia gravis | Neurological conditions | 1 | 1 | | | y | | | | ever |
| 224 | Parkinson's disease | Neurological conditions | 1 | 1 | | | y | | | | ever |
| 225 | Peripheral neuropathies (excluding cranial nerve and carpal tunnel syndromes) | Neurological conditions | 1 | 1 | | | y | | | | ever |
| 226 | Postviral fatigue syndrome, neurasthenia and fibromyalgia | Neurological conditions | 1 | 1 | | | y | | | | ever |
| 227 | Trigeminal neuralgia | Neurological conditions | 1 | 1 | | | y | | | | ever |
| 228 | Acne | Skin conditions | 1 | 1 | | | y | | | | ever |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 229 | Actinic keratosis | Skin conditions | 1 | 1 | | | y | | | | ever |
| 230 | Alopecia areata | Skin conditions | 1 | 1 | | | y | | | | ever |
| 231 | Dermatitis (atopic/contact/other /unspecified) | Skin conditions | 1 | 1 | | | y | | | | ever |
| 232 | Hidradenitis suppurativa | Skin conditions | 1 | 1 | | | y | | | | ever |
| 233 | Lichen planus | Skin conditions | 1 | 1 | | | y | | | | ever |
| 234 | Pilonidal cyst/sinus | Skin conditions | 1 | 1 | | 1 | y | | | | ever |
| 235 | Psoriasis | Skin conditions | 1 | 1 | | | y | | | | ever |
| 236 | Rosacea | Skin conditions | 1 | 1 | | | y | | | | ever |
| 237 | Seborrheic dermatitis | Skin conditions | 1 | 1 | | | y | | | | ever |
| 238 | Urticaria | Skin conditions | 1 | 1 | | | y | | | | ever |
| 239 | Vitiligo | Skin conditions | 1 | 1 | | | y | | | | ever |

### 10.6.4 List of excluded conditions

This table is available online at https://github.com/rmjlrwh/FatigueRiskMap

| Disease name | Disease group | Primary code list source was Kuan et al | Exclusion reason |
|---|---|---|---|
| Adrenal adenoma | Diseases of the Endocrine System | | Code list unavailable |
| Osteonecrosis | Musculoskeletal conditions | | Code list unavailable |
| Paget's Disease | Musculoskeletal conditions | | Code list unavailable |
| Toxoplasmosis | Infectious Diseases | | Code list unavailable |
| Biliary duct strictures | Diseases of the Digestive System | | Code list unavailable |
| Vitamin D deficiency | Diseases of the Endocrine System | | Available codes do not capture fully; need prescriptions and test results too |
| Post cricoid web | Diseases of the Digestive System | | Code list unavailable |
| Thrombocytosis | Haematological/Immunological conditions | | Symptom, not disease |
| Pancreatic duct dilatation | Diseases of the Digestive System | | Code list unavailable |
| Hydrosalpinx | Diseases of the Genitourinary system | | Code list unavailable |
| Plantar fasciitis | Musculoskeletal conditions | | Code list unavailable |
| Small bowel stricture | Diseases of the Digestive System | | Code list unavailable |
| Vertebral fractures | Musculoskeletal conditions | | Code list unavailable |
| Pelvic cysts | Diseases of the Genitourinary system | | Code list unavailable |
| Gastric intestinal metaplasia | Diseases of the Digestive System | | Code list unavailable |
| Liver lesions | Benign Neoplasm/CIN | | Code list unavailable |
| EBV | Infectious Diseases | | Code list unavailable |
| Cauda equina/cord compression | Musculoskeletal conditions | | Code list unavailable |
| Sacroillitis | Musculoskeletal conditions | | Code list unavailable |

| | | | |
|---|---|---|---|
| Benign/indeterminate lung changes | Benign Neoplasm/CIN | | Code list unavailable |
| Fractured neck of femur | Musculoskeletal conditions | | Code list unavailable |
| Para-thyroid adenoma | Diseases of the Endocrine System | | Code list unavailable |
| Hip labral tear | Musculoskeletal conditions | | Code list unavailable |
| Acromegaly | Diseases of the Endocrine System | | Code list unavailable |
| Cushing's syndrome | Diseases of the Endocrine System | | Code list unavailable |
| Hypopituitarism | Diseases of the Endocrine System | | Code list unavailable |
| Vasculitis/CT disorders | Musculoskeletal conditions | | Code list unavailable |
| Heavy metal toxicity | Haematological/Immunological conditions | | Code list unavailable |
| Appendix mucocele | Diseases of the Digestive System | | Code list unavailable |
| Q fever | Infectious Diseases | | Code list unavailable |
| Haemochromatosis | Haematological/Immunological conditions | | Code list unavailable |
| Mesenteric panniculitis | Diseases of the Digestive System | | Code list unavailable |
| Brucellosis | Infectious Diseases | | Code list unavailable |
| Gallbladder polyps | Diseases of the Digestive System | | Code list unavailable |
| Addison’s disease | Diseases of the Endocrine System | | Code list unavailable |
| Thoracic outlet syndrome | Diseases of the Circulatory System | | Code list unavailable |
| Benign bony lesions | Benign Neoplasm/CIN | | Code list unavailable |
| Raised SFLC/paraproteins | Haematological/Immunological conditions | | Code list unavailable |
| Pancreatic cysts/lesions | Benign Neoplasm/CIN | | Code list unavailable |
| Myometrium adenomyosis | Diseases of the Genitourinary system | | Code list unavailable |
| Cytomegalovirus | Infectious Diseases | | Code list unavailable |
| Secondary pulmonary hypertension | Diseases of the Circulatory System | y | Not incident disease |
| Cystic fibrosis | Diseases of the Endocrine System | y | Age-irrelevant (e.g. usually diagnosed in childhood) |
| Obesity | Diseases of the Endocrine System | y | Symptom, not disease |
| Postcoital and contact bleeding | Diseases of the Genitourinary system | y | Symptom, not disease |
| Postmenopausal bleeding | Diseases of the Genitourinary system | y | Symptom, not disease |

| | | | |
|---|---|---|---|
| Undescended testicle | Diseases of the Genitourinary system | y | Age-irrelevant (e.g. usually diagnosed in childhood) |
| Pulmonary collapse (excl pneumothorax) | Diseases of the Respiratory System | y | Symptom, not disease |
| Secondary polycythaemia | Haematological/Immunological conditions | y | Not incident disease |
| Hyposplenism | Haematological/Immunological conditions | y | Age-irrelevant (e.g. usually diagnosed in childhood) |
| Primary or Idiopathic Thrombocytopaenia | Haematological/Immunological conditions | y | Symptom, not disease |
| Secondary or other Thrombocytopaenia | Haematological/Immunological conditions | y | Symptom, not disease |
| Sickle-cell trait | Haematological/Immunological conditions | y | Symptom, not disease |
| Splenomegaly | Haematological/Immunological conditions | y | Symptom, not disease |
| Thalassaemia trait | Haematological/Immunological conditions | y | Symptom, not disease |
| Thrombophilia | Haematological/Immunological conditions | y | Symptom, not disease |
| Intellectual disability | Mental Health Disorders | y | Age-irrelevant (e.g. usually diagnosed in childhood) |
| Juvenile arthritis | Musculoskeletal conditions | y | Age-irrelevant (e.g. usually diagnosed in childhood) |
| Cerebral Palsy | Neurological conditions | y | Age-irrelevant (e.g. usually diagnosed in childhood) |
| Diabetic neurological complications | Neurological conditions | y | Not incident disease |
| Congenital malformations of cardiac septa | Perinatal conditions | y | Age-irrelevant (e.g. usually diagnosed in childhood) |
| Down's syndrome | Perinatal conditions | y | Age-irrelevant (e.g. usually diagnosed in childhood) |
| High birth weight | Perinatal conditions | y | Age-irrelevant (e.g. usually diagnosed in childhood) |
| Intrauterine hypoxia | Perinatal conditions | y | Age-irrelevant (e.g. usually diagnosed in childhood) |
| Slow fetal growth or low birth weight | Perinatal conditions | y | Age-irrelevant (e.g. usually diagnosed in childhood) |
| Neonatal jaundice (excl haemolytic dz of the newborn) | Perinatal conditions | y | Age-irrelevant (e.g. usually diagnosed in childhood) |
| Patent ductus arteriosus | Perinatal conditions | y | Age-irrelevant (e.g. usually diagnosed in childhood) |
| Post-term infant | Perinatal conditions | y | Age-irrelevant (e.g. usually diagnosed in childhood) |
| Prematurity | Perinatal conditions | y | Age-irrelevant (e.g. usually diagnosed in childhood) |
| Respiratory distress of newborn | Perinatal conditions | y | Age-irrelevant (e.g. usually diagnosed in childhood) |
| Bacterial sepsis of newborn | Perinatal conditions | y | Age-irrelevant (e.g. usually diagnosed in childhood) |

| | | | |
|---|---|---|---|
| Spina bifida | Perinatal conditions | y | Age-irrelevant (e.g. usually diagnosed in childhood) |

### 10.6.5 Phenotypes used to define each condition

This table is available online at https://github.com/rmjlrwh/FatigueRiskMap

Preview below:

| readcode | readterm | medcode | disease_new | disease_number_new |
|---|---|---|---|---|
| M261A00 | Pustular acne | 4360 | Acne | 228 |
| M261B00 | Steroid acne | 32041 | Acne | 228 |
| M261z00 | Other acne NOS | 15655 | Acne | 228 |
| M261012 | Comedo | 15609 | Acne | 228 |
| M261C00 | Tropical acne | 21182 | Acne | 228 |
| M261F00 | Acne fulminans | 53994 | Acne | 228 |
| M261X00 | Acne, unspecified | 25590 | Acne | 228 |
| M261J00 | Acne necrotica | 34937 | Acne | 228 |
| 2FG5.00 | Acne scar | 17895 | Acne | 228 |
| M261400 | Iodine acne | 67453 | Acne | 228 |
| M260000 | Acne frontalis | 10183 | Acne | 228 |
| M261011 | Blackhead | 8065 | Acne | 228 |
| M26y200 | Giant comedo | 9058 | Acne | 228 |
| Myu6800 | [X]Other acne | 52909 | Acne | 228 |
| N25..00 | SAPHO syndrome Synov, Acne | 25737 | Acne | 228 |
| Myu6F00 | [X]Acne, unspecified | 33367 | Acne | 228 |
| M261.00 | Other acne | 4065 | Acne | 228 |
| M261600 | Cystic acne | 1654 | Acne | 228 |
| M261900 | Occupational acne | 31828 | Acne | 228 |
| M261G00 | Acne agminata | 43811 | Acne | 228 |
| M260z11 | Acne necrotica | 20164 | Acne | 228 |
| M261300 | Chlorine acne | 55983 | Acne | 228 |

### 10.6.6 Cohort size and disease risk by previous diagnoses

Sensitivity analysis: Cohort size, unadjusted 1-year risk and rank of incident diseases in men and women aged 30-99 years, by cohort (fatigue presenters, non-fatigue presenters, registered patients), before and after removing patients with a previous diagnosis of each disease.

Results for groups with under <6 diagnosed patients are suppressed to reduce disclosivity risk.

This table is available online at https://github.com/rmjlrwh/FatigueRiskMap

Preview below:

| Sex | Cohort | Disease group no. | Disease group | Disease no. | Disease | No previous disease cohort | No previous disease | No previous disease model | No previous disease 95% | No previous disease 95% | All patients : cohort (N) | All patients : case (n) | All patients - modelled risk | All patients - 95% lower bound | All patients - 95% upper bound | Absolute difference in risk (%) if including | 95% CI comparison between patients with no previous |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Men | Fatigue presenters | 3 | Diseases of the Cardiovascular system | 25 | Hypertension | 69636 | 3546 | 5.092 | 4.931 | 5.258 | 100881 | 13178 | 13.063 | 12.856 | 13.272 | 7.971 | Higher (significant) |
| Women | Fatigue presenters | 3 | Diseases of the Cardiovascular system | 25 | Hypertension | 157892 | 4835 | 3.062 | 2.978 | 3.148 | 204033 | 17873 | 9.76 | 9.638 | 9.883 | 5.698 | Higher (significant) |
| Men | Fatigue presenters | 3 | Diseases of the Cardiovascular system | 22 | Coronary heart disease not otherwise specified | 88629 | 1795 | 2.025 | 1.935 | 2.12 | 100881 | 5645 | 5.596 | 5.456 | 5.739 | 3.571 | Higher (significant) |
| Women | Fatigue presenters | 7 | Diseases of the Endocrine System | 87 | Hypo or hyperthyroidism | 176955 | 4297 | 2.428 | 2.358 | 2.501 | 204033 | 11841 | 5.803 | 5.703 | 5.906 | 3.375 | Higher (significant) |
| Women | Fatigue presenters | 10 | Diseases of the Respiratory System | 124 | Asthma | 176416 | 1495 | 0.847 | 0.806 | 0.891 | 204033 | 8516 | 4.174 | 4.088 | 4.261 | 3.327 | Higher (significant) |
| Women | Fatigue presenters | 12 | Infectious Diseases | 177 | Urinary Tract Infections | 173805 | 11572 | 6.658 | 6.542 | 6.776 | 204033 | 20115 | 9.859 | 9.73 | 9.989 | 3.201 | Higher (significant) |
| Women | Fatigue presenters | 13 | Mental Health Disorders | 185 | Depression | 147470 | 5364 | 3.637 | 3.543 | 3.734 | 204033 | 13396 | 6.566 | 6.459 | 6.674 | 2.929 | Higher (significant) |
| Men | Fatigue presenters | 3 | Diseases of the Cardiovascular system | 17 | Atrial fibrillation | 93621 | 1770 | 1.891 | 1.805 | 1.98 | 100881 | 4516 | 4.477 | 4.351 | 4.606 | 2.586 | Higher (significant) |
| Men | Fatigue presenters | 10 | Diseases of the Respiratory System | 124 | Asthma | 89797 | 576 | 0.641 | 0.591 | 0.696 | 100881 | 3195 | 3.167 | 3.061 | 3.277 | 2.526 | Higher (significant) |
| Men | Fatigue presenters | 12 | Infectious Diseases | 167 | Lower Respiratory Tract Infections | 88937 | 4933 | 5.547 | 5.398 | 5.699 | 100881 | 7982 | 7.912 | 7.747 | 8.08 | 2.365 | Higher (significant) |
| Men | Fatigue presenters | 9 | Diseases of the Genitourinary system | 103 | Chronic kidney disease | 90855 | 1922 | 2.115 | 2.024 | 2.211 | 100881 | 4402 | 4.364 | 4.239 | 4.491 | 2.249 | Higher (significant) |
| Women | Fatigue presenters | 12 | Infectious Diseases | 153 | Ear and Upper Respiratory Tract Infections | 170371 | 11314 | 6.641 | 6.524 | 6.76 | 204033 | 18050 | 8.847 | 8.724 | 8.971 | 2.206 | Higher (significant) |
| Women | Fatigue presenters | 12 | Infectious Diseases | 167 | Lower Respiratory Tract Infections | 181181 | 8685 | 4.794 | 4.696 | 4.893 | 204033 | 13936 | 6.83 | 6.722 | 6.941 | 2.036 | Higher (significant) |
| Men | Fatigue presenters | 13 | Mental Health Disorders | 185 | Depression | 82951 | 2659 | 3.209 | 3.092 | 3.332 | 100881 | 5134 | 5.089 | 4.955 | 5.227 | 1.88 | Higher (significant) |
| Men | Fatigue presenters | 4 | Diseases of the Circulatory System | 42 | Stable angina | 93005 | 974 | 1.047 | 0.984 | 1.115 | 100881 | 2623 | 2.6 | 2.504 | 2.7 | 1.553 | Higher (significant) |

### 10.6.7 Age-adjusted risk

*1-year risk and rank of incident** diseases in men and women aged 30-99 years, by cohort (fatigue-presenters, non-fatigue presenters, registered patients), after adjusting for age.*

Patients with previous diagnoses of each disease are excluded, so the cohort for each disease is different. See appendices for relevant sensitivity analyses. We show expected risk in non-fatigue presenters and registered patients, if the age profiles were the same as for fatigue-presenters. Diseases with < 100 diagnosed patients in fatigue presenters are excluded, and results for groups with under <6 diagnosed patients are suppressed to reduce disclosivity risk.

This table is available online at [https://github.com/rmjlrwh/FatigueRiskMap](https://github.com/rmjlrwh/FatigueRiskMap)

Preview below:

| Sex | Disease group no | Disease group | Disease no | Disease | Fatigue presenters: cases (n) | Fatigue presenters: modelled risk | Fatigue presenters: 95% lower bound | Fatigue presenters: 95% upper bound | Fatigue presenters: rank | Non-fatigue presenters: cases (n) | Non-fatigue presenters: model | Non-fatigue presenters: 95% lower | Non-fatigue presenters: 95% upper | Non-fatigue presenters: rank | Registered patients: cases (n) | Registered patients: modelled risk | Registered patients: 95% lower | Registered patients: 95% upper | Registered patients rank | Absolute excess risk (%) in fatigue vs non-fatigue | 95% CI comparison between fatigue & non-fatigue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Women | 12 | Infectious D | 177 | Urinary Tract Infections | 11572 | 6.66 | 6.54 | 6.78 | 1 | 7683 | 4.42 | 4.32 | 4.52 | 2 | 6970 | 4.01 | 3.92 | 4.1 | 1 | 2.24 | Higher (significant |
| Women | 12 | Infectious D | 153 | Ear and Upper Respiratory Tract Infections | 11314 | 6.64 | 6.52 | 6.76 | 2 | 7744 | 4.54 | 4.45 | 4.64 | 1 | 6271 | 3.68 | 3.59 | 3.77 | 2 | 2.1 | Higher (significant |
| Women | 12 | Infectious D | 167 | Lower Respiratory Tract Infections | 8685 | 4.79 | 4.7 | 4.89 | 3 | 5933 | 3.28 | 3.19 | 3.36 | 3 | 5305 | 2.93 | 2.85 | 3.01 | 3 | 1.52 | Higher (significant |
| Women | 13 | Mental Hea | 195 | Depression | 5364 | 3.64 | 3.54 | 3.73 | 4 | 3883 | 1.28 | 1.22 | 1.33 | 7 | 3879 | 1.27 | 1.22 | 1.33 | 7 | 2.36 | Higher (significant |
| Women | 3 | Diseases of | 25 | Hypertension | 4835 | 3.06 | 2.98 | 3.15 | 5 | 3098 | 1.96 | 1.9 | 2.03 | 4 | 3275 | 2.07 | 2 | 2.15 | 4 | 1.1 | Higher (significant |
| Women | 14 | Musculoske | 195 | Connective & soft tissue disorders | 4300 | 2.73 | 2.65 | 2.81 | 6 | 2820 | 1.79 | 1.72 | 1.85 | 5 | 2506 | 1.59 | 1.53 | 1.65 | 5 | 0.94 | Higher (significant |
| Women | 7 | Diseases of | 87 | Hypo or hyperthyroidism | 4297 | 2.43 | 2.36 | 2.5 | 7 | 1181 | 0.67 | 0.63 | 0.71 | 21 | 1171 | 0.66 | 0.62 | 0.7 | 21 | 1.76 | Higher (significant |
| Women | 13 | Mental Hea | 187 | Insomnia & sleep disturbances | 3863 | 2.25 | 2.18 | 2.32 | 8 | 1364 | 0.8 | 0.75 | 0.84 | 16 | 1387 | 0.81 | 0.77 | 0.85 | 16 | 1.46 | Higher (significant |
| Women | 16 | Skin conditi | 231 | Dermatitis [atopic/contact/other/unspecifie | 3367 | 2.15 | 2.08 | 2.22 | 9 | 2216 | 1.42 | 1.36 | 1.48 | 6 | 2013 | 1.29 | 1.23 | 1.34 | 6 | 0.73 | Higher (significant |
| Women | 13 | Mental Hea | 181 | Anxiety disorders | 3442 | 2.08 | 2.01 | 2.15 | 10 | 1600 | 0.97 | 0.92 | 1.01 | 12 | 1547 | 0.93 | 0.89 | 0.98 | 13 | 1.11 | Higher (significant |
| Women | 9 | Diseases of | 103 | Chronic kidney disease | 3231 | 1.73 | 1.67 | 1.79 | 11 | 1464 | 0.78 | 0.74 | 0.82 | 17 | 1872 | 1 | 0.96 | 1.05 | 11 | 0.95 | Higher (significant |
| Women | 9 | Diseases of | 114 | Menorrhagia and polymenorrhoea | 2960 | 1.73 | 1.67 | 1.79 | 12 | 1317 | 0.77 | 0.73 | 0.81 | 18 | 1218 | 0.71 | 0.67 | 0.75 | 19 | 0.96 | Higher (significant |
| Women | 12 | Infectious D | 151 | Bacterial Diseases (excl TB) | 3381 | 1.69 | 1.64 | 1.75 | 13 | 2104 | 1.05 | 1.01 | 1.1 | 8 | 2480 | 1.24 | 1.19 | 1.29 | 8 | 0.64 | Higher (significant |
| Women | 14 | Musculoske | 204 | Osteoarthritis (excl spine) | 2751 | 1.61 | 1.55 | 1.67 | 14 | 1778 | 1.04 | 1 | 1.09 | 9 | 1688 | 1 | 0.95 | 1.04 | 12 | 0.57 | Higher (significant |
| Women | 11 | Haematolog | 142 | Other anaemias | 2860 | 1.52 | 1.47 | 1.58 | 15 | 897 | 0.48 | 0.45 | 0.51 | 31 | 948 | 0.5 | 0.47 | 0.54 | 29 | 1.05 | Higher (significant |
| Women | 5 | Diseases of | 69 | Gastro-oesophageal reflux disease | 2722 | 1.51 | 1.45 | 1.57 | 16 | 1507 | 0.84 | 0.79 | 0.88 | 14 | 1343 | 0.74 | 0.71 | 0.78 | 18 | 0.67 | Higher (significant |
| Women | 12 | Infectious D | 172 | Other or unspecified infectious organisms | 3013 | 1.5 | 1.45 | 1.56 | 17 | 2006 | 1 | 0.96 | 1.05 | 10 | 2289 | 1.14 | 1.1 | 1.19 | 9 | 0.5 | Higher (significant |

### 10.6.8 Age-specific risk

*Age-specific modelled 1-year risk and rank of incident diseases in men and women presenting with new-onset fatigue, compared to non-fatigue presenters and registered patients, by year of age*

Patients with previous diagnoses of each disease are excluded, so the cohort for each disease is different. See appendices for relevant sensitivity analyses. Diseases with < 100 diagnosed patients in fatigue presenters are excluded, and results for groups with under <6 diagnosed patients are suppressed to reduce disclosivity risk.

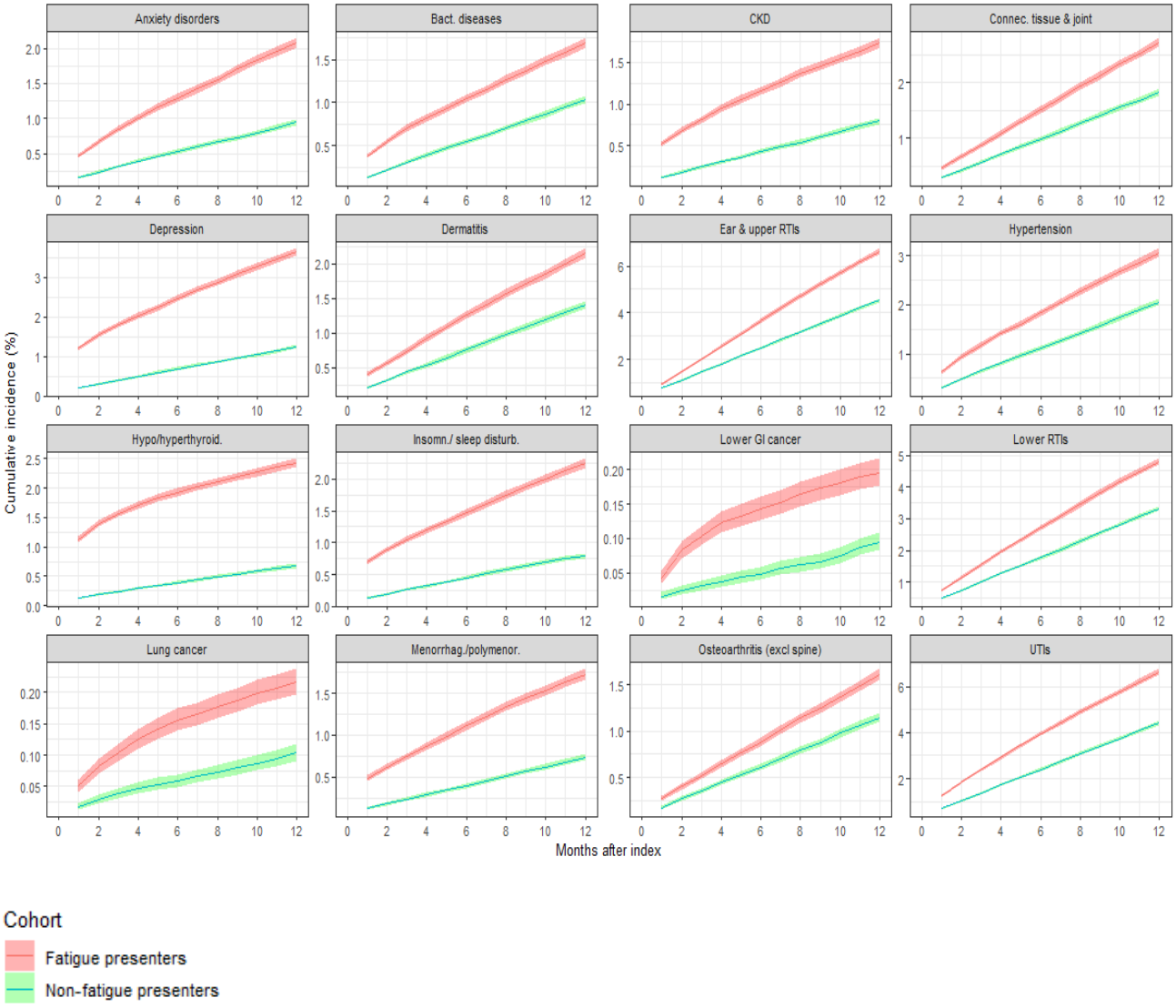
This table is available online at https://github.com/rmjlrwh/FatigueRiskMap

Preview below:

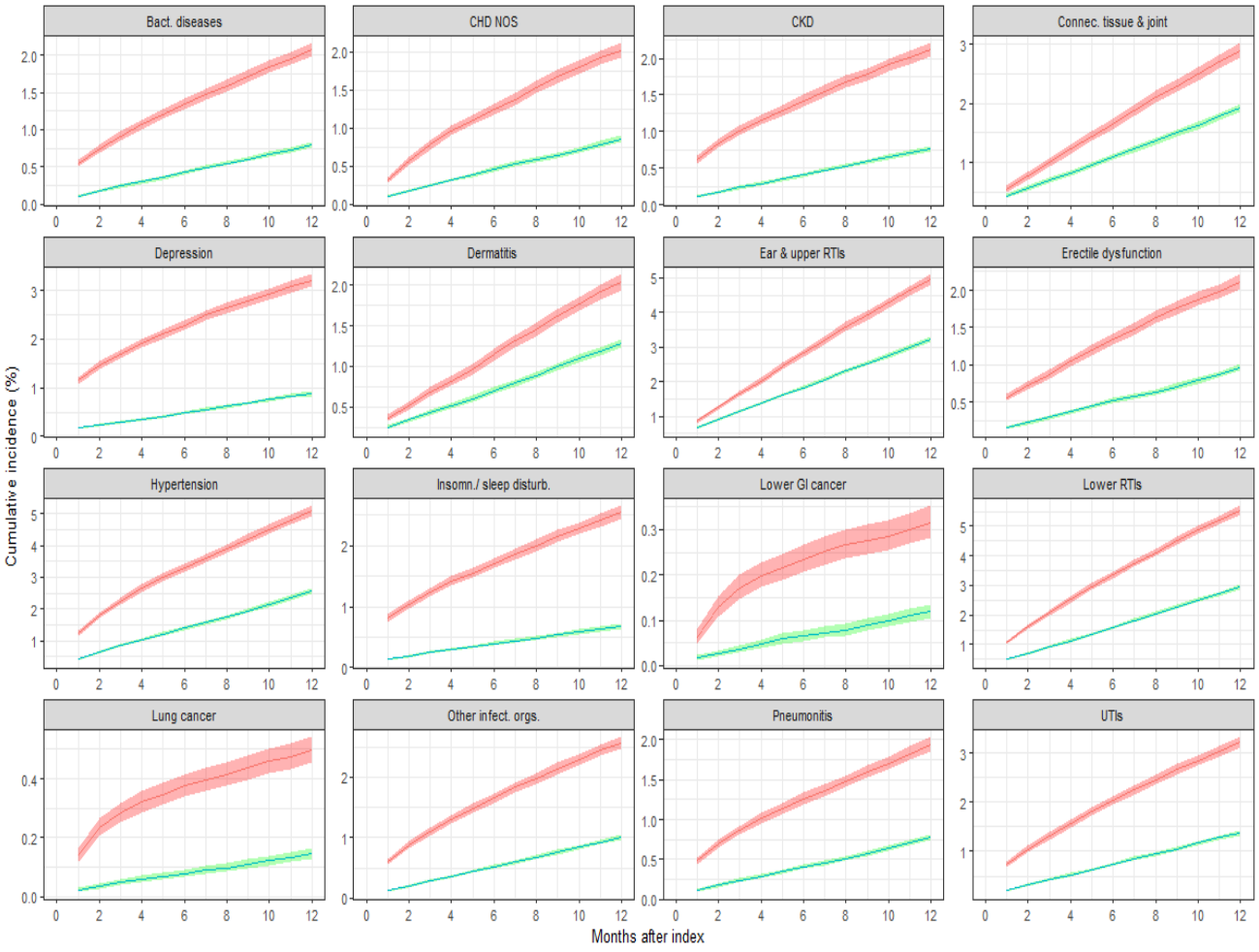
| Se | Age ge | Disea se gro ne | Disea se gro | Dis se i | Disease | Fatigue present ers: modelle d risk (%) | Fatigue present ers: 95% lower boun (%) | Fatigue present ers: 95% upper boun (%) | Non-fatigue presenters: modelled risk (%) | Non-fatigue presenters: 95% lowe bound (% | Non-fatigue presenters: 95% uppe bound (% | Registered patients: modelled risk (%) | Registered patients: 95% lowe bound (% | Registered patients: 95% uppe bound (% | Absolute excess risk (%) in fatigue vs non-fatig presente | Disea ran | 95% CI comparison between fatigue & non-fatigue presenter |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Women | 60 | 13 | Mental H | 185 | Depression | 3.44 | 3.24 | 3.65 | 1.06 | 0.97 | 1.15 | 1.11 | 1.04 | 1.19 | 2.38 | 1 | Higher (significant) |
| Women | 60 | 12 | Infectiou | 153 | Ear and Upper Respiratory Tract Infect | 6.85 | 6.59 | 7.12 | 4.59 | 4.41 | 4.78 | 3.87 | 3.74 | 4.02 | 2.26 | 2 | Higher (significant) |
| Women | 60 | 7 | Disease | 87 | Hypo or hyperthyroidism | 2.88 | 2.71 | 3.07 | 0.7 | 0.62 | 0.77 | 0.72 | 0.66 | 0.79 | 2.19 | 3 | Higher (significant) |
| Women | 60 | 12 | Infectiou | 177 | Urinary Tract Infections | 5.94 | 5.7 | 6.19 | 3.76 | 3.59 | 3.93 | 3.38 | 3.25 | 3.52 | 2.18 | 4 | Higher (significant) |
| Women | 60 | 12 | Infectiou | 167 | Lower Respiratory Tract Infections | 5.01 | 4.78 | 5.24 | 3.13 | 2.97 | 3.29 | 2.91 | 2.79 | 3.04 | 1.88 | 5 | Higher (significant) |
| Women | 60 | 13 | Mental H | 187 | Insomnia & sleep disturbances | 2.54 | 2.38 | 2.71 | 0.85 | 0.77 | 0.93 | 0.82 | 0.76 | 0.89 | 1.69 | 6 | Higher (significant) |
| Women | 60 | 3 | Disease | 25 | Hypertension | 3.7 | 3.48 | 3.93 | 2.1 | 1.96 | 2.26 | 2.37 | 2.24 | 2.5 | 1.6 | 7 | Higher (significant) |
| Women | 60 | 14 | Musculo | 195 | Connective & soft tissue disorders | 3.68 | 3.47 | 3.9 | 2.27 | 2.13 | 2.42 | 2.04 | 1.93 | 2.15 | 1.41 | 8 | Higher (significant) |
| Women | 60 | 14 | Musculo | 204 | Osteoarthritis (excl spine) | 3.01 | 2.8 | 3.22 | 1.78 | 1.65 | 1.92 | 1.65 | 1.55 | 1.77 | 1.23 | 9 | Higher (significant) |
| Women | 60 | 13 | Mental H | 181 | Anxiety disorders | 1.96 | 1.82 | 2.12 | 0.79 | 0.71 | 0.87 | 0.78 | 0.72 | 0.84 | 1.18 | 10 | Higher (significant) |
| Women | 60 | 5 | Disease | 69 | Gastro-oesophageal reflux disease | 1.96 | 1.82 | 2.11 | 1.06 | 0.97 | 1.16 | 0.97 | 0.9 | 1.04 | 0.9 | 11 | Higher (significant) |
| Women | 60 | 9 | Disease | 103 | Chronic kidney disease | 1.46 | 1.33 | 1.6 | 0.59 | 0.53 | 0.67 | 0.67 | 0.61 | 0.74 | 0.86 | 12 | Higher (significant) |
| Women | 60 | 16 | Skin cor | 231 | Dermatitis (atopic/contact/other/unspe | 2.13 | 1.98 | 2.3 | 1.36 | 1.25 | 1.48 | 1.24 | 1.16 | 1.33 | 0.77 | 13 | Higher (significant) |
| Women | 60 | 7 | Disease | 85 | Diabetes | 1.33 | 1.21 | 1.46 | 0.8 | 0.53 | 0.68 | 0.65 | 0.59 | 0.71 | 0.73 | 14 | Higher (significant) |
| Women | 60 | 9 | Disease | 119 | Urinary Incontinence | 1.23 | 1.12 | 1.34 | 0.55 | 0.49 | 0.62 | 0.54 | 0.49 | 0.6 | 0.67 | 15 | Higher (significant) |
| Women | 60 | 2 | Cancers | 1001 | All cancers combined | 1.58 | 1.45 | 1.72 | 0.91 | 0.82 | 1 | 1.12 | 1.04 | 1.2 | 0.67 | 16 | Higher (significant) |
| Women | 60 | 10 | Disease | 121 | Allergic and chronic rhinitis | 1.24 | 1.13 | 1.36 | 0.58 | 0.52 | 0.65 | 0.57 | 0.52 | 0.63 | 0.66 | 17 | Higher (significant) |
| Women | 60 | 11 | Haemato | 142 | Other anaemias | 0.89 | 0.8 | 0.98 | 0.25 | 0.21 | 0.29 | 0.26 | 0.23 | 0.3 | 0.64 | 18 | Higher (significant) |
| Women | 60 | 11 | Haemato | 141 | Iron deficiency anaemia | 0.84 | 0.77 | 0.92 | 0.21 | 0.17 | 0.24 | 0.2 | 0.17 | 0.23 | 0.64 | 19 | Higher (significant) |

### 10.6.9 Monthly cumulative risk

Men



Women


Bact. diseases
CHD NOS
CKD
Connec. tissue & joint
Depression
Dermatitis
Ear & upper RTIs
Erectile dysfunction
Hypertension
Insomn./ sleep disturb.
Lower GI cancer
Lower RTIs
Lung cancer
Other infect. orgs.
Pneumonitis
UTIs
Cumulative incidence (%)
Months after index
Cohort
Fatigue presenters
Non-fatigue presenters