INSTITUTE OF CLINICAL TRIALS AND METHODOLOGY

Utilising electronic health records (EHR) to improve the efficiency of oncology clinical trials

Dr Archibald James Cameron Macnair

20/07/2022

A thesis submitted for the degree of Doctor of Medicine by Research

UCL

Declaration

I, Dr Archibald James Cameron Macnair 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.

……

Dr Archibald James Cameron Macnair

Abstract

Introduction

Randomised controlled trials (RCTs) are typically considered the gold standard for investigating oncology treatments. However, RCTs can be resource intensive for healthcare organisations. Electronic health records (EHRs) comprising routinely collected data have the potential to improve the efficiency of clinical trials by supporting recruitment, follow-up and lowering costs. However, there is limited information about the reliability (accuracy and completeness) of EHR and how they can be used to support clinical trials in the UK.

Methods

The utility of EHR data was examined by: 1) comparing trial data relating to cardiovascular serious adverse events with data from two EHR databases using sensitivity and predictive value. 2) assessing the practicality and validity of data from a national cancer registry as a passive method of long-term follow-up in an established RCT and 3) whether data on aspirin use from a pre-existing RCT dataset, could provide further supporting evidence for the initiation of a new trial.

Results

Access to national EHRs was challenging with time from application to receiving data ranging from 8-15 months. Some cardiovascular inpatient events i.e. acute coronary syndrome matched well to trial data (sensitivity 0.89). However, both the trial data and EHR found separate events meaning poor comparability within these datasets. Long-term passive follow-up with EHR could be used for mortality data (sensitivity 100%) and potentially cancer recurrence if multiple cancer registry datasets were used. A large trial evaluating aspirin supported by EHR is feasible, but this initial study did not provide statistically supportive evidence for the use of aspirin in primary chemoprevention.

Conclusion

EHR have the potential, in the future, to support clinical trials particularly for long-term follow-up. However there remain major hurdles to resolve prior to their widespread use

especially in access to appropriate datasets and understanding the limitations of the data.

Impact statement

This body of work contributes to the continued assessment of EHR in relation to clinical trials. EHR are continuing to become more widespread, and permit linkage between datasets, but are not designed or set up for clinical trial use. Some clinical trials have used them despite a lack or limited evidence to demonstrate parity to traditional trial data collection. This study demonstrates that depending on the outcome measure EHR could be used to follow-up participants in some circumstances potentially reducing trial costs, decreasing loss to follow-up and improving data integrity. However, they must be used appropriately with transparent integration into trial protocols from the start and clarity in how they were used when results are published. The evidence from this body of work could contribute to the discussion with trial regulators about how EHR can be used in the future and the appropriate development of procedures to allow successful integration.

During the time I worked on this thesis I helped colleagues to advocate for better access to these databases for researchers. During the covid crisis the ability to access up to date data from national databases radically improved but only for covid specific projects. The efficient access for academic research needs to be maintained. My publication on the difficulties of access and working with database holders was cited in the National Cancer Research institute (NCRI) response to the recent Goldacre review of health data use in clinical trials. Holders of datasets also should consider the needs of clinical trialists to allow for appropriate retention of the data in contracts which is economically sustainable.

Lastly this body of evidence demonstrates to policy makers and governmental bodies the importance and potential of these national datasets. These resources need to be appropriately funded and linked so that they can continue to support the four nations to work together to improve UK academia.

Contributions

This methodology work involved several different trials and also was integrated into the electronic health record methodology work stream within the MRC CTU. Many different people have been involved in helping with the projects from all the trial teams and the methodology specialist group. I led and wrote the data applications to all various national datasets. Data application process also involved continued work with the datasets to define data needed and verify data governance for the transfer of data. I also led writing the paper on the experience of accessing data which was published in Trials journal. I also designed the methodology comparison studies using Add-Aspirin and PATCH trial data. I wrote and designed the protocol for the primary prevention study with aspirin using UKCTOCS EHR data. During my time working with the methodology group on EHR I also helped with a review of EHR use in the UK and helped write a paper on the subject. I also collaborated on the formation of a protocol for a trial within a trial in the comparing death data with EHR and trial data.

Acknowledgements

Firstly, I would like to thank my two supervisors, Ruth Langley and Duncan Gilbert, for allowing me the opportunity to work with them and the team at MRC CTU, and for their supervision and guidance, from which I have learnt a great deal. I am grateful to the MRC CTU at UCL, without whose funding this research would not have been possible. I would like to thank the EHR methodology group for the help throughout this research and all of the trial teams. I would like to particularly thank the trial statisticians Matthew Nankivell, Mathew Burnell and Fay Cafferty for their statistical support for each of the projects involving their trials. I would like to thank the other clinical fellows in the unit Nal, Hannah and Bhav for their continued support and friendship. I would especially like to thank my wife Catriona and my son Angus and daughter Eliza for their continued patience and for all the fun and laughter in my breaks from writing. Lastly to all the participants in the trials whose EHR made this study possible.

Contents

Table of Tables

Table of Figures

Abbreviations

ACS	Acute Coronary Syndrome
A&E	Accident and Emergency
APC	Admitted Patient Care
BMI	Body Mass Index
CAG	Confidentiality Advisory Group
CCGs	Clinical Commissioning Groups
CHD	Coronary Heart Disease
CI	Confidence Interval
COURAGE	Clinical Outcomes Utilizing Revascularization and Aggressive Drug Evaluation trial
COPD	Chronic obstructive pulmonary disease
COSD	Cancer outcomes and services data set
COX	Cyclo-oxygenase enzyme
CPRD	Clinical Practice Research Datalink
CRF	Case Report Form
CRUK	Cancer Research UK
CTCAE	Common Terminology Criteria for Adverse Events
CTU	Clinical trials unit
CVA	Cerebrovascular accident
CVS	Cardiovascular
CWT	Cancer Waiting Times dataset
DARG	Data Access Request Group
DARS	Data Access Request Service
DCIS	Ductal carcinoma in situ
DIDs	Diagnostic Imaging Data Set
DVT	Deep vein thrombosis
ECG	Electrocardiogram
EHR	Electronic Health Record
EMR	Electronic Medical Record
EPOCH	Enhanced Peri-Operative Care for High-risk patients
ER	Oestrogen receptor
FDA	US Food and Drug Administration
GCP	Good Clinical Practice
GDPR	General Data Protection Regulation
GI	Gastrointestinal
GP	General Practitioner
HR	Hazard Ratio
HES	Hospital Episode Statistics
HIV	Human Immunodeficiency Virus
HPFS	Health Professionals Follow-Up Study
HQIP	Health Quality Improvement Partnership
HSCIC	Health and Social Care Information Centre
ICD-10	International Classification of Diseases, Tenth Revision
IDMC	Independent Data Monitoring Committee
LHRHa	Luteinising Hormone Releasing Hormone agonists

MDT	Multi-Disciplinary Team
MHMDS, MHLDDS, MHSDS	Mental Health data
MHRA	Medicines and Healthcare products Regulatory Agency
MI	Myocardial infarction
MINAP	Myocardial Ischaemia National Audit Project
MRC	Medical Research Council
NCMP	National Child Measurement Programme
NCRAS	National Cancer Registration Service
NCRN	National Cancer Research Network
NDA	National Diabetes Audit
NDRS	National Disease Registration Service
NHFA	National Heart Failure Audit
NHS	National Health Service
NICE	National Institute for Health Care and Excellence
NICOR	National Institute for Cardiovascular Outcomes Research
NIHR	National Institute of Health Research
NOS	Not otherwise specified
ODR	Office of Data Release
ONS	Office of National Statistics
OPCS	Operating Procedure Codes Supplement
PATCH	Prostate Adenocarcinoma TransCutaneous Hormone trial
PBCR	Population-Based Cancer Registries
PCI	Percutaneous coronary intervention
PDS	Personal Demographics Service
PE	Pulmonary Embolism
PHE	Public Health England
PLCO	Prostate, Lung, Colorectal and Ovarian Screening Study
PPV	Positive Predictive Value
PROMs	Patient Reported Outcome Measures
PSA	Prostate specific antigen
QALY	Quality-adjusted life year
REC	Research Ethics Committee
RCT	Randomised control trial
RR	Relative Risk
RTDS	Radiotherapy dataset
SAE	Serious adverse event
SACT	Systemic anti-cancer treatment
SCARR	Swedish Coronary Angiography and Angioplasty Registry
SHIFT	Self Harm Intervention: Family Therapy study
STEMI	ST segment elevation myocardial infarction
SUS PbR	Secondary Uses Service Payment by results
SWAT	Study within a trial
TACT2	Trial of accelerated adjuvant chemotherapy with capecitabine in early breast cancer
TASTE	Thrombus Aspiration in ST-Elevation Myocardial Infarction study
tE2	Transdermal oestrogen patches
TIA	Transient ischaemic attack

TSC	Trials steering committee
UCL	University College London
UK	United Kingdom
UKCTOCS	UK Collaborative Trial of Ovarian Cancer Screening
USPSTF	US Preventative services task force
WHO	World Health Organisation
WOSCOPS	West of Scotland coronary prevention study

1 Introduction

1.1 Global cancer burden

Cancer has been described as the ‘Emperor of all maladies’ in a recent book describing the history of cancer due to the biology of the disease and societies’ perception that medicine is unable to cure the disease (1). However, modern medicine has meant that many cancers have become curable or at least treatable with some cancers now described as a chronic disease. Yet, these advances in treatment have not been reflected in the global burden of the disease with cancer incidence continuing to increase annually. In 2020 there was an estimated 19.3 million new cancer cases and 10 million deaths (2). Many cancers, especially advanced disease on presentation, are still not curable and the disease is expected to rank as the leading cause of death and the single most important barrier to increasing life expectancy in every country in the 21st century (2). The reason for increasing incidence is multi factorial but the most important factors are considered to be increasing age and a westernisation of lifestyles.

The growing incidence and prevalence of cancer in the United Kingdom (UK) population reflects the global trend. In the UK there were on average 363,000 new cancer cases every year between 2014-2016 and incidence rates have increased by 13% since the 1990s (3). The incidence of cancer in the UK is ranked higher then two thirds of Europe and higher than 90% of the world (3). This along with mental health are the top priorities for the National Health Service (NHS) set out in their 5 year forward plan (4, 5). Fundamentally in the long-term the factors that cause cancer have to be addressed such as obesity rates, alcohol consumption and smoking tobacco through lifestyle change (4). There is an equally important role in improving the earlier diagnosis of cancer and developing chemoprevention agents. The treatment of cancer must also improve to achieve higher cure rates and to improve the prognosis of patients with advanced incurable disease. Better therapies will emerge from translational science and by demonstrating efficiency with clinical trials.

1.2 Clinical trials and cancer

Despite the long history of cancer treatment, it was not until the 1940's that clinical trials played a major part in the development of new treatments. One of the first clinical trials was performed by Farber and demonstrated one of the first remissions of paediatric leukaemia (1). Since then many, if not most, improvements in cancer outcomes have been based upon the results of clinical trials. These trials have evaluated prevention and screening strategies, diagnosis, treatment, and the reduction of side effects from cancer treatments.

Cancer is cured predominately by surgery and radiotherapy, however, the bulk of cancer clinical trials have been based on the development of new cancer drugs (6, 7). The most highly regarded and influential type of trial/ study is the randomized control trial (RCT). A RCT is "a trial in which subjects are randomly assigned to one of two groups: one (the experimental group) receiving the intervention that is being tested, and the other (the comparison group or control) receiving the alternative (conventional treatment). The groups are then followed up to see if there are any differences between them in outcome." (8). The importance of RCTs versus other clinical research is that other studies cannot rule out the possibility that the association was caused by a third factor linked to both the intervention and outcome (9). Well designed RCTs are considered the gold standard of comparative studies (10).

Over the last 20 years the rate of new drugs licensed for cancer treatment has continued to rise with a particular increase over the last 10 years (11). This is illustrated by the US Food and Drug Administration (FDA) drug approvals where 25% have been for the treatment of cancer in 2018 (12). These have all been based solely on RCTs. However there are some challenges to RCTs with high costs, volunteer bias and ethical dilemmas at times.

There are some that argue that it is not ethical that patients are given a placebo instead of an active treatment in RCTs (13). Also the time it takes for RCTs to accrue participants, analyse and publish data may delay the introduction of potentially lifesaving treatment. A further challenge is that medical practice may have moved on by the time the results of a trial are published, for example in the COURAGE trial (Clinical Outcomes Utilizing Revascularization and Aggressive Drug Evaluation; NCT

00007657) where newer drug eluting stents were already in use by the time the investigators published less efficacious results with bare metal stents. At times the perceived delays in RCT processes enter public consciousness as was the case in the Human Immunodeficiency Virus (HIV) community in the 1980's, who demanded access to the new medications before the trials had finished due to the rising epidemic of the disease (13).

Cancer trials are funded by different types of organisations e.g. charitable, governmental, pharmaceutical companies and large international consortia. Drug trials are mostly funded by the pharmaceutical companies themselves where academic trials may need funding from one or all of the categories described above. Cancer Research UK (CRUK) has been one of the largest charitable funders of cancer research in the UK and demonstrates the huge cost as they spent £546 million from 2018-2019 on all types of cancer research (14) and one of the governmental funders, National Institute of Health Research (NIHR), spending £31.5 million on cancer research alone (15). These large sums of money reflect the cost of the randomized control drug trials which dominate cancer research. One of these trials can cost millions to run (16). The main contributors to this cost are salary costs, patient enrolment, treatment, and the follow-up phase (16). As with many complex initiatives the cost of the administration can be the most significant (11-29%) and the cost of monitoring the trial for a regulatory body can also dominate a budget (9-14%) (17). During the planning and initiation of any clinical trial, cost is a key consideration and often deters funders from supporting a more diverse array of topics as they need to prioritise based on limited resource. Cost can also make funders risk adverse as the financial implication of funding an unsuccessful project can be significant especially to the pharmaceutical industry. Importantly, given the primary outcome measures relevant to obtaining definitive answers may require long-term patient data, this also increase costs. This can mean follow-up is not long enough to collect all the potential important effects from a treatment (18).

Another challenge of RCTs is the risk of bias towards a particular population. This can be due to the exclusion or inclusion criteria of the group studied not reflecting the general population (19). Equally many RCTs only recruit in high income countries such that treatment regimens may only be relevant to those countries and not globally

generalizable (20). Biases also occur in follow-up. For example patients with cancer progression maybe less likely to attend follow-up depending on the setting of the trial as they have deteriorated medically. There are also concerns that certain vulnerable groups like elderly patients are more likely to be lost to follow-up giving unintentional bias to studies (21-23).

In summary, RCTs need greater diversity of participants throughout the trial, and lower costs. Cancer trials, to be sustainable in the future, must address these problems. Improving trial methodology or using new technologies may provide solutions.

1.3 Electronic health records

The definition of electronic health records (EHR) on first review is relatively simple. Broadly defined- 'EHRs represent longitudinal data (in electronic format) that are collected during routine delivery of health care' (24). This describes EHR as a whole and is an area that has grown considerably as computing technology developed through the latter part of the 20th century. As health informatics have matured so has the use of electronic records to facilitate day-to-day healthcare but they are now also being considered as tools to improve clinical research. The fundamental parts of the above definition are that these are records that are 'electronic' and are not developed for the sole purpose of research. New electronic methods of collecting case report forms or complex databases to store and refine study data will not be reviewed.

The broad term of EHR can have different meanings depending upon the interaction with the data. For example, Kimberley *et al.* in their review differentiate between electronic health records (EHR) and electronic medical records (EMR) as follows:

'EHRs are electronic platforms that contain health-related data collected during medical care in practices, clinics and other medical settings from various sources, connected to form a network of patient clinical data. EHRs can also incorporate software that allow straightforward physician ordering practice (CPOEs), even including safety features, or that guide physicians through clinical decision-making with up-to-date guidelines (CDSS).' (25).

'EMRs are routinely collected data sources that contain standard medical and clinical data gathered during medical care in an individual location of a practice, clinic or other medical setting. When the data are shared among different locations and units, it becomes a network and is considered an EHR (i.e., an electronic chart system in a primary care practice that cannot be accessed by any other entity is an EMR, whereas a hospital system in which laboratory data, affiliated clinic charts, etc., are all accessed under 1 platform is an EHR).' (25).

The distinction between the two is fundamental as EHR encompass more than a single general practitioner (GP) practice recording their data electronically but includes a network or a platform that has a collection of multiple different clinical records.

The applicability of EHR are considerable providing clinical support tools to healthcare professionals, electronic ordering and prescribing, telehealth and opportunities for patient reported outcomes measures capture (25). All of these applications can help clinicians with patient care and the administrative aspects of medicine. EHR can also capture clinical activity for health performance management. These applications are also being developed to assist clinical trials in the identification of potential participants and for trial follow-up and have been shown to reduce costs and streamline trial design (26-29).

1.4 Evolution of electronic health record - registry data

Another form of EHR are national databases, audits and registries that routinely collect data from their countries of origin and map to important healthcare events. Collection of data within health care systems globally is not new and has been collected for several decades e.g. cancer registries. However, the ability to store this data electronically and link national datasets has evolved to a standard that is now considered potentially useable for research. The NHS has a long history of collecting data and with all patients having a designated NHS number, this can act as a consistent identifier within datasets. Other countries have national registries with varying degree of data quality and reporting based on their healthcare system. This is most clearly seen within the World Health Organisation (WHO) 2019 report of monitoring health for the sustainable development goals. Primary data availability from

member states included in the report is highly variable. A third of the countries do not have underlying data for this report especially in low and middle income countries. Mortality data is crucial for monitoring sustainable goals but only half of the countries are able to register 80% of adult deaths (30).

Of particular interest are the cancer registries for oncology trials. A review of global cancer incidence suggests that only 1 in 3 cancer registries globally have high quality data and only 1 in 5 report equivalent mortality data to the WHO (31). Even in higher income countries in Europe there is not universal coverage of cancer registration (32). Often national statistics on incidence are based on regional registries which report to the European network of cancer registries (32). In Ferley's *et al.* publication on European cancer incidence and cancer mortality, Greek cancer incidence had to be derived from neighbouring countries (32). This suggests that data coverage may not be currently sufficient to support clinical trials. Equally, comparisons between countries the data has to be coded in a similar fashion. Again this is not universal and even in Europe, with a strong cancer registry network, the standard of data can be different from country to country and even region (32). As registries are country specific in structure and coverage I will focus on the UK and English registries that our particularly relevant to oncology trials.

The national health data registries within the UK are devolved. Therefore England, Scotland, Wales, and Northern Ireland hold their own national registries. The major registries are Hospital Episodes Statistics (HES), national disease registries and national audits. All these registries are controlled and facilitated by different governmental organisations depending on the devolved nation. Each country has their own rules for data access.

1.4.1 Hospital Episode Statistics (HES)

HES data is probably the most widely requested and used for multiple purposes often by local authorities and Clinical Commissioning Groups (CCGs) (33). HES is controlled by NHS Digital within England. In Scotland this is called the Information Service Division; Wales it is named Patient Episode Database for Wales; and Northern Ireland it has the title of Hospital Activity Statistics and is controlled by the Department of

Health. All of these datasets have comparable data but can differ slightly depending on the country of origin. All intend to gather information on all secondary care activity that includes accident and emergency (A&E) attendance, hospital inpatient admission and outpatient appointments.

NHS Digital in England has been the custodian of HES data since 2016. Prior to this it was the Health and Social Care Information Centre (HSCIC) from 2005. Prior to 2005 it was part of the department of Health and NHS information authority. The data recorded from inpatient admission or Admitted Patient Care (APC) has been recorded since 1989, with outpatient data recorded from 2003 and A&E attendance since 2007. A major change to the administration and also data quality followed the Health and Social Care act in 2012 where there was a drive to improve data within the NHS (34, 35). Information on various attributes of the patient care can be assessed with more than 500 variables ranging from diagnosis, admission dates and times to discharge date and demographics of the patient and provider (36).

The HES data is primarily a resource for the reimbursement of hospital activity. At present inpatient admissions have the greatest variability in reimbursement depending upon the diagnosis, activity, and duration of the stay and therefore is the greatest financial driver for a hospital (37). For example hospitals will get reimbursed significantly more for a patient requiring an organ transplant compared to an admission for a community acquired pneumonia with minimal comorbidities. Outpatient and A&E settings have a significantly higher volume of patients but do not require as many data fields to create the cost reimbursement. These submissions are completed manually and certain more complex data fields such as diagnosis are often poorly populated (37, 38). Therefore, due to the quality and completeness of data the APC dataset is often thought to be the most reliable for research but does depend on the research question.

NHS Digital within England are now able to provide access to many other different datasets a selection is in **Table 1.1**. This is an evolving list as NHS Digital are continually adding further datasets. One of the key strengths is that many of these datasets can be linked to each other based on patient identifiers and also to other national datasets such as cancer registry data or national clinical audits.

Table 1-1: Datasets held by NHS Digital (39)

Datasets	Dataset description
Hospital Episodes Statistics (HES)	Over 1 billion records of patients attending accident and emergency units, admitted for treatment or attending outpatient clinics at NHS hospitals in England.
Patient Reported Outcome Measures (PROMs)	Pre and post-operative survey data collected from patients receiving hip replacement, knee replacement, hernia and varicose vein surgery.
Diagnostic Imaging Data Set (DID)	Data on NHS-funded diagnostic imaging tests, such as scans and x-rays, extracted from NHS providers' radiological information systems.
Mental Health data (MHMDS, MHLDDS, MHSDS)	Includes adults in receipt of NHS funded specialist, secondary mental health or learning disability services.
NHS registration Data from the Personal Demographics Service	The Personal Demographics Service (PDS) is the national electronic database of NHS patient demographic details such as name, address, date of birth, NHS Number and registered GP.
Civil Registrations (Deaths)- Secondary Care Cut	Information including the date, place and cause of death from the Office for National Statistics (ONS).
Secondary Uses Service Payment by results (SUS PbR)	National secondary care payment information from the Secondary Uses Service.
Adult Psychiatric Morbidity Survey	Provides data on the prevalence of both treated and untreated psychiatric

	disorder in the English adult population (aged 16 and over).
National Child Measurement Programme (NCMP)	The program that measures the height and weight of children in Reception class (aged 4 to 5) and year 6 (aged 10 to 11), to assess overweight and obesity levels in children within primary schools. It is recognized internationally as a world-class source of public health intelligence and holds UK National Statistics status.
National Diabetes Audit (NDA) Core	One of the largest annual clinical audits in the world. It measures the effectiveness of diabetes healthcare against NICE Clinical Guidelines and NICE Quality Standards in England and Wales, for both primary and secondary care.

To access any data from NHS Digital all enquiries proceed through the Data Access Request Service (DARS) via their online portal. Each data access application is considered depending on their lawful basis, data security infrastructure, technical feasibility and purpose including benefit to health and social care in the UK (40). There is a fee for data applications depending on the number of datasets requested, linkage, number of years of data, frequency of data download required and length of retention of the data. Researchers are increasingly using this system and NHS Digital are creating NHS DigiTrials which is a service which will specifically help trialists to help with applications, feasibility of trials and long-term data for trial/ study use (41).

1.4.2 National Cancer Registration Service (NCRAS)

For those involved in the design and management of oncology clinical trials potentially the most significant and important EHR are the cancer registries. Population-Based Cancer Registries (PBCRs) aim to identify all cases of cancer that occur in a defined population and collect a minimum number of 10 variables for each case (42). All cancer registries within the UK are devolved but are linked by the UK and Ireland association of cancer registries (43). Within England, Public Health England (PHE) have managed the cancer registry since 2013 under the name of National Cancer Registration Service (NCRAS), now at time of writing, under the umbrella term of National Disease Registration Service (NDRS) (44). Though the cancer registry has been established since the 1980's, it was not, until the Calman-Hine report in 1995 and the development of Multi-Disciplinary Team (MDT) meetings with registration systems for cancer waiting times did data quality significantly improve (44). Since then these datasets have been extended to include information on cancer stage, pathology, systemic anti-cancer treatment (SACT) and radiotherapy. NCRAS have the ability to link various datasets that are either held by them or NHS Digital/ Office of National Statistics (ONS). NCRAS integrate data from patient administration systems, MDT meetings systems, pathology and imaging reports to form their cancer registry. They also have feeds from other data holders like NHS Digital with datasets such as Cancer waiting times, HES and DID and ONS mortality data. The key datasets within NCRAS form from all this data are described in **Table 1.2**.

Table 1-2: NCRAS datasets and description (45)

Dataset	Description
Cancer outcomes and services data set (COSD)/ Cancer registration	National standard for reporting in the NHS. This dataset includes data on the patient, their diagnosis, tumour and treatment events.
Cancer Waiting times (CWT)	This dataset supports the management and monitoring of waiting times for the full course of the cancer diagnostic and treatment pathway. This provides information on time to referrals, first definitive treatment and also subsequent treatments for their primary cancer and cancer recurrences. This information is provided by NHS Digital.
National Radiotherapy (RTDS)	This dataset includes radiotherapy episodes for a patient including information on preparation, planning and delivery of treatment as covered by their treatment intention.
Systemic Anti-Cancer Therapy (SACT)	This dataset collects clinical management on patients receiving cancer chemotherapy in or funded by NHS in England. It includes all settings of delivery. This includes all drug treatments with an anti-cancer effect including traditional cytotoxic chemotherapy and all newer agents.
National Cancer Audits	Tumour type specific audits to compare the diagnostic and treatment pathways, and outcomes for those diagnosed in England and Wales

At the time of writing the control of data access for this registry was via ODR. The same principles for data access to HES via DARS is also used by the ODR. The team within the ODR work with analysts from PHE to verify if data requests are achievable and how best to use the datasets. In the future the cancer registry is likely to come under the umbrella of NHS Digital as the data controller so access may also be through the NHS DARS system.

1.4.3 Health Quality Improvement Partnership (HQIP) National Audits and National Institute for Cardiovascular Outcomes Research (NICOR)

There are a number of national clinical audits that are established within the UK. Health Quality Improvement Partnership (HQIP) oversees and governs many of these audits and currently describes 28 of across a number of disease areas examples include the National Adult Diabetes Audit and the Myocardial Ischaemia National Audit Project (MINAP) (46). In a recent review of all these audit datasets many govern themselves outside the remit of HQIP (33). HQIP are often the data controller for the audits and therefore hold a similar role to the ODR within NCRAS considering if the application is valid. The process for data access can be more complicated with initial applications having to be agreed by the audit team first and then the request reviewed by the HQIP Data Access Request Group (DARG). Applicants can be charged by both the audit organisation and then HQIP separately.

Of particular relevance to this thesis is the audits held by National Institute for Cardiovascular Outcomes Research (NICOR). NICOR collects routine EHR data from hospital trusts within England and Wales. NICOR was originally hosted by University College London (UCL) but moved to Barts Health NHS Trust in 2017. They produce reports to enable hospitals and healthcare improvement bodies to monitor and improve the care and outcomes of patients with cardiovascular disease. NICOR manages six national clinical audits and a number of new health technology registries including the National Heart Failure Audit (NHFA), MINAP, Adult Cardiac Surgery (Surgery Audit), Adult Percutaneous Coronary Interventions (Angioplasty Audit), Cardiac Rhythm Management (Arrhythmia Audit), and Congenital Heart Disease in Children and Adults (Congenital Audit) (47). The data is collected directly from

hospitals and also recently linked to HES data to verify the data. The data collected in NICOR is more detailed than routine datasets such as HES.

1.4.4 Clinical Practice Research Datalink (CPRD)

There are also a number of datasets specifically related to primary health care. At present there is no one data set for all primary practices in the UK or England. However, there are datasets that have been designed primarily for research in primary care. Clinical Practice Research Datalink (CPRD) is an example. It is a network of GP practices across the UK and potentially encompasses 35 million patients (48). It has access to data recorded by GP practices including diagnosis and prescription data. The coding of this data is dependent upon the electronic medical record that the GP practice uses, and can be linked to other national datasets. At present CPRD has generated over 2,200 peer-reviewed publications including both retrospective and prospective studies (48).

1.5 Benefits and challenges of EHR within clinical trials

1.5.1 Benefits of EHR with clinical trials

The use of routinely collected data within clinical trials has been described as ‘the next disruptive technology for clinical trials’ (49). The author suggests that it has the potential to transform existing standards, procedures and cost structures within clinical trials (49). With any disruptive technology there are potential benefits but also challenges which could mean that implementation is difficult.

The most exciting potential benefit of EHR is the opportunity to create pragmatic registry based RCTs. This is defined as ‘pragmatic trials that use registries as a platform for case records, data collection, randomization and follow-up’ (50). The advocates for pragmatic trials describe the negative attributes of traditional RCTs with high costs, patient population bias and loss to follow-up. They argue that with the use of registries, trialists can help solve the ‘catch 22’ situation of not doing randomized trials into common therapies as they would not be financially viable or too difficult to

do due to regulation that was designed to protect the patient (10). The registry-based trial therefore aims to include the statistical rigor of traditional RCTs and also expedites and enhances patient enrolment, minimises cost and addresses the generalisability of findings (51).

This is best demonstrated by the Thrombus Aspiration in ST-Elevation Myocardial Infarction (TASTE) (NCT 01093404) trial which was based upon Swedish registries (52). They used a novel trial design that evaluated routine intracoronary thrombus aspiration before percutaneous coronary intervention (PCI) in patients with ST segment elevation myocardial infarction (STEMI) versus PCI alone. It was a multicentre prospective randomized controlled open label clinical trial. The trial used the national comprehensive Swedish Coronary Angiography and Angioplasty Registry (SCARR) for identification of patients, randomization, collection of baseline and procedural variables and follow-up. The participants verbally gave consent on admission with STEMI and were then randomly allocated 1:1 within the registry by an online module to thrombus aspiration prior to angioplasty +/- stent or angioplasty +/- stent alone. The participants confirmed their consent to participate in writing within 24 hours. The trial enrolled 7,244 participants and demonstrated that a technique that was widely used in clinical practice did not reduce rate of death (Hazard Ratio (HR), 0.94; Confidence interval (CI) 0.72-1.22; p=0.63), hospitalization for recurrent myocardial infarction at 30 days (HR, 0.61; CI 0.34-1.07; p=0.09) or stent thrombosis at 1 year (HR, 0.47; CI 0.20 to 1.02; p=0.06) (52).

The trial not only evaluated a clinically important question but also demonstrated that the use of registries for all aspect of the study was feasible. Within 2 years and 9 months 61% patients of all patients who presented with a diagnosis of a STEMI and were referred for PCI in Sweden were enrolled into the trial representing 82% of all eligible patients. The main reason why patients were not enrolled were they could not give verbal consent as they were unconscious at presentation. No patients were lost to follow-up (52). The cost of the trial was low at $50 per patient or $300,000 in total for the whole trial (49).

Cancer registries have often been used in other types of clinical studies. For example there is a long history of their use in observational studies which were subsequently published in high impact journals (53-57). One exemplar is a publication from Herrett

et al. and co-workers in the Lancet detailing a retrospective cohort study where 1,222,670 patients were followed for a median of 4.3 years using CPRD, HES and ONS data. The authors demonstrated that a cardiovascular risk-based strategy (QRISK2 ≥10%) could prevent over a third more cardiovascular disease events than the 2011 NICE cardiovascular guideline and a fifth more than the 2019 NICE cardiovascular guideline.

One of the most appealing indications for registry use in terms of clinical trials is for long-term follow-up. RCTs often have funding for a designated time which may not completely align to the outcome of the intervention that is being investigated (58). A study in hormone sensitive breast cancer may need 10-20 years follow-up to see if patients have recurrence based on the biology of the cancer (59). This may not be logistically or financially viable. Also the loss to follow-up rate within trials can be around 5% (59). EHR could conceivably collect this information reducing administrative support and costs. This has been demonstrated successfully in a number of trials including cancer trials (21, 24, 60).

As an example, the West of Scotland coronary prevention study (WOSCOPS) evaluated the effect of pravastatin (versus placebo) on death from coronary heart disease or definite non-fatal myocardial infarction, long-term follow-up after the first primary analysis was conducted using EHR only (61, 62). Data included GP prescription data, hospital discharge records, cancer registry and general register office death records which could be linked via the Information and Statistical Division of NHS Scotland. The data showed continued benefit at 10 years for those who remained on pravastatin with a reduction in death from coronary heart disease or non-fatal myocardial infarction from 15.5% to 11.8% (HR 0.73; 95% CI 0.63-0.83; p=<0.001). The risk reduction from the primary analysis reduced from 40% to 18% after long-term follow (61).

Another potential benefit of EHR could be the verification of follow-up data. This could be either verification of source data from sites or from patient self-reported events depending upon the design of the trial. Data from registries can also be used as a trigger to go back to site/ clinician to see if an event has actually occurred and to obtain more details. This has the benefit of strengthening the validity of the results by cross-checking with another source of data especially in self-reported outcomes from

questionnaires. It also has the advantage that as national registries have national coverage then participants are less likely to be lost to follow-up. This 'trigger' method could provide an efficient and more cost effective way of following up patients. It could also reduce workload at trial sites as they would not have to routinely contact participants but only arrange a follow-up visit when an outcome occurred in the registry data. This approach is used in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS) (NCT00058032) (60). This was a RCT of 202,638 post-menopausal women between the ages of 50-74 who either received annual multimodal screening, annual trans-vaginal ultrasound or no screening. The way that this trial was carried out was completely dependent on EHR. Several sources were used to see if participants had received a diagnosis of ovarian cancer. If the registries highlighted a patient with cancer then the trial team would go back to relevant NHS provider to collect the relevant information. The national registries were also used to validate other sources of data particularly a self-reported cancer diagnosis via questionnaires (60). This was a large and long running screening trial and EHR were one of the reasons why this trial could be run in a cost effective manner. The multiple data sources provided reassurance all cancers were being identified minimal input from the healthcare professionals involved in their care.

Additionally EHR can be one of the most effective ways of demonstrating the cost of a new intervention. Many national registries are designed to facilitate cost recovery either from health insurance companies or for national reimbursement such as in the NHS. Therefore, this can give a potentially more realistic cost of healthcare use within trials. EHR can demonstrate that an intervention can save the healthcare system money in the future. A good example of this is demonstrated in the Self Harm Intervention: Family Therapy (SHIFT) trial (ISRCTN59793150) (63). This was a pragmatic phase 3 multi-centre RCT of evaluating systemic family therapy treatment versus treatment as usual for young people after self-harm. Overall this trial demonstrated that systemic family therapy treatment conferred no added benefit. They were also able to demonstrate through a cost effective analysis using HES data that the cost of this intervention per Quality-adjusted life year (QALY) was above the National Institute for Health Care and Excellence (NICE) threshold.

1.5.2 Challenges to the use of EHR within clinical trials

Despite the potential benefits for the use of EHR in clinical trials use of registry data is still only used by a minority of clinical trials. This was demonstrated in a review of data access requests to UK registries where approximately only 3% of registered clinical trials used registry data (33). This is in contrast to the number of clinical trials stating they hope to use EHR data in their trial which has grown significantly (64). This is probably due to the fact that there are ongoing concerns about the utility (reliability, completeness, and accuracy) of EHR, and accessibility of the data.

Cost reduction is thought to be one of the major benefits of using EHR as described previously in the TASTE and WOSCOPS trials. However, when reviewed in more detail the cost of retrieving data from the biggest EHR providers can be relatively expensive. The cost of data is not usually calculated on the number of participants that you request data for but rather on the number of linked databases, the duration of data needed and also the number of regular extracts needed (65). For NHS digital at the time of writing, quarterly downloads of HES APC and mortality data over 5-years costs approximately £90,000 excluding value-added tax (65). For large trials, this may be relatively cost-effective if this was one of the sole sources of follow-up for a large cohort of participants. However, for a small trial this may not be feasible. There is also an additional cost of keeping the data which can often be incurred when contracts run out or if you want to extend the period of time that the data is kept within the research group. If a trial has to hold data, due to trial regulations, for a certain time after the trial has closed then these costs can continue for many years.

One of the greatest concerns for all that use EHR is how complete or reliable is the data? The answer seems to be that is very dependent on the data source and is especially true of registries across different countries. In this respect, the UK, USA and Scandinavians countries are thought to have good coverage and data reliability (66-71). In Scandinavia there has been high investment and a long legacy of multiple data registries for numerous conditions with procedures in place for high quality data collection and validation of data sets. In other high income countries many have difficulties with incomplete population coverage due to insufficient funding (72). If clinical trials were based across several countries the reliability of each data set would have to be scrutinised and validated to see if each countries' data was of a sufficiently

high standard for the outcome required (73). Comparing data between countries may be difficult as not all may use the same diagnosis or procedure codes (74).

In the UK, the major registries mostly have complete population coverage, data quality and linkage due to good quality procedures and audit (75, 76). For trials though errors in either dataset linkage or data error could have a significant consequence on the result of the trial. One trial noted that there was a failure of the registry to update their patient list with the new patients recruited into the trial. The trial team therefore only received mortality data on half of their cohort (77). Another study demonstrated that linkage between data sets were not as complete as expected quoting a missed match of 4.1% and also a false match of 0.2% between HES and a paediatric intensive care registry (78). Both these processes have improved over the years and reports on validity and data integrity are in the public domain via the websites of the various datasets.

Registries in the UK are not designed for research and there is still concern that data is not appropriate for clinical trials. This is partly due to the way the data is collected particularly HES which is the most used data source in the UK. HES data is collected by each provider and centralised first to a secondary user service and then passed to NHS digital to be cleaned. For some variables like diagnosis or type of operation there is hospital level variation that produces inconsistencies in coding as site clinical coders interpret the medical notes differently. These inconsistencies can produce mistakes and an inappropriate code for that event. Registries often get information for the same patient from many different sources and they must interpret and judge what the correct information is. This can also lead to inappropriate information for individual patients being recorded. There is evidence that when registries are compared for the same outcome there can be a significant degree of inconsistency (79). Not all data in clinical trials is correct and random error can occur which are unlikely to change the result of a large trial. However if there is fundamental systemic errors in the registry this could change the result (80).

In addition as registries need to collect information from multiple different sources there can be a significant time lag which would be inappropriate for trials particularly in reference to adverse reactions to trial treatment. This has improved over time and HES data in the UK now can only have delays in order of weeks. However, cancer

registries in the UK have a delay of approximately 18 months which is not dissimilar to other countries (71, 75). Audit data, such as NICOR, is often only given for the latest audit publication which again could encompass a lag period of a couple of years. The registries need this time to make sure that the correct data is published.

Data access to registries have become increasing more complex and difficult over the years. In Europe this is partly due to the new General Data Protection Regulation (GDPR) rules which in the UK resulted in the 2018 Data Protection Act (81). Additionally registries have become stricter on access based on the purpose of the project and benefit to society. This results from previous inappropriate releases of data and the subsequent Partridge review (82). This has meant that the application for these datasets have become increasingly complicated to adhere to data law.

The data security infrastructure to support transfer and storage of data must also be extremely high. The cost and resource needed for this security means that only large institutions such as universities have the capability to hold the data. Registries have also had exacting standards for consent. Wording previously approved by trial regulators or ethics committee has been deemed not sufficient. The advice on the wording of consent can also change with time which means trials can have appropriate consent at conception but not subsequently during the course of the trial. These strict laws and data security often have been too complex causing significant delays in retrieving the data or sometimes not at all (83, 84). Several trials have had difficulty publishing results due to data access (33). One example is the Enhanced Peri-Operative Care for High-risk patients (EPOCH) trial (ISRCTN80682973), where the research team were unable to procure mortality data in Wales following hospital admissions. As a result the researchers had to change their planned primary analysis to make sure their publication was not delayed significantly (85).

As with all valid research the raw data of the trial is often shared with other researchers or with meta-analysis groups (86). Registries have put in place strict rules about sharing of registry data with other research groups to the extent that trials using registry data may not be able to continue this practice (84). If registry data is going to be used more widely in trials the issue relating to data access/ holding of data and subsequent sharing will need to be addressed.

There is also concern from the regulators of the trials themselves. According to Good Clinical Practice (GCP), data has to have an obvious audit trail with oversight and visibility of the data collection and processing (87). This becomes difficult if using registry data as it is cleaned and handled by many different organisations and personnel before it is given to the trial team. This is because it is initially entered into hospital records by clinical staff, then coded locally before sending it nationally to be cleaned and validated. The final process could be reviewed, however, the processing at local hospitals would be very challenging. The Medicines and Healthcare products Regulatory Agency (MRHA) are working with the trial community to establish guidance on using real world data in the form of EMRs and registries. At the time of writing this was still in a consultation phase awaiting final approval (88).

There are huge potential benefits of using EHR in clinical trials increasing the scope of conditions and treatments that can be reviewed; decreasing the cost; improving long-term follow-up and recruiting a broader cross section of society. These are all the things that RCTs are criticised for not addressing and in the appropriate setting registry data could hold the answer. This ‘disruptive technology’ needs improved regulation and data laws to allow appropriate and timely data access and for the storing and sharing of data in accordance with trial regulation. There also needs to be more methodological work to make sure that the data is accurate, reliable and useful to the academic community and trial regulators.

1.6 Aims

The overarching question for this thesis is ‘how can EHR be used to optimise oncology clinical trials’. The thesis will use certain active trials to answer various questions on trial design which pertain to the use of EHR in oncology trials. For each trial a specific question will be evaluated using different national registry data. The aim would be to demonstrate the current utility of registry data for these trials and also for future trials currently in the design phase. This will be achieved looking at the following aspects of clinical trial methodology.

1. Can some serious adverse events be collected for oncology trials using EHR in registries and audits? This will be evaluated in the Prostate Adenocarcinoma

TransCutaneous Hormone trial (PATCH) trial (ISRCTN70406718) where participants with hormone sensitive prostate cancer were randomly allocated to transdermal oestradiol patches (tE2) or Luteinising hormone-releasing hormone agonists (LHRHa). Cardiovascular disease is a secondary outcome measure. Cardiovascular serious adverse events will therefore be reviewed through a triangulation of data with trial data, NHS Digital HES data and NICOR audit data to assess if electronic health records are consistent with trial data.

2. Can EHR be employed in the long-term follow-up patients participating in oncology trials? This will be assessed within the Add-Aspirin trial (NCT02804815) where participants are randomly allocated to aspirin or placebo for at least 5 years following potentially curable therapy for early stage malignancies. The trial design includes a long-term analysis of overall survival 15 years post randomization which will encompass oncological outcomes from the original malignancy, adverse effects of aspirin and the development of second malignancies. EHR data will be accessed to see whether it can be used for long-term follow-up to provide relevant outcome data, and if the data currently available is consistent with the data collected directly from the participating sites.
3. Consider the use of electronic health records in a primary prevention trial of cancer? As part of the thesis a concept of using aspirin as a chemo-preventative agent in primary prevention setting in high risk patients was devised. As preliminary work for this proposed trial a study was designed to add to the current evidence of aspirin in the primary prevention setting and also to test current EHR capability for this future trial. This ethically approved study was designed within the UKCTOCS trial, a screening trial in ovarian cancer. This will review if aspirin can decrease the risk of cancer incidence in a cohort of women in the UKCTOCS trial and will use electronic heath records to collect the necessary outcome data for a future trial in the aspirin primary prevention setting.

2 Comparison of cardiovascular serious adverse events in the PATCH trial with national registry data and audit data

2.1 Introduction

The primary outcome of RCTs typically describes the efficacy of a drug or intervention on the disease which is being studied. In oncology trials this is often measured in length of overall survival of the patient with a certain type of cancer between groups or how long the cancer takes to progress in each comparison. A key role, however, is also to understand the safety profile of the intervention (along with additional information such as e.g. costs of treatment) in order that treatment decisions that are in the best interest of both patient and society can be made.

Central to this is the assessment of adverse events, defined as data that is recorded in clinic by clinicians on direct questioning of the participant to collect unfavourable and unintended signs, symptoms or disease that is temporally associated with the use of a drug (87). Trial regulations stipulated by GCP and overseen by regulators mandates that the reporting of *serious* adverse events (SAE) (those that result in hospitalisation or are life threatening/result in patient death) is expedited and subject to enhanced reporting, such that that responsible bodies within the trial team (e.g. independent data monitoring committee (IDMC)) are alerted to evidence of patient harm from the experimental approach. The collection and reporting of this data is essential for any RCT and is monitored throughout the course of the trial by the trial physicians and IDMC. The process for reporting these adverse events has traditionally been done on paper case report forms (CRF) that the site sends into the sponsor either after a follow-up appointment or in the case of a SAE then 24 hours from the site's knowledge of the event.

2.1.1 Challenges of traditional adverse event reporting

Adverse events collection and publication in trials then is essential to allow physicians to have detailed conversations with patients about the risk/ benefit of starting a new treatment. The highest quality data for adverse events is still thought to be that collected during clinical trials. Historically however, adverse events have generally been poorly reported within trials (89) and in cancer trials specifically (90, 91). This could be either at the site as it can be time consuming for clinicians, or system failures at site in collection under report certain adverse events. One example of how this might occur is when a patient attends a different hospital to that where they are usually treated with a SAE. If the patient/treating hospital team never informs their site trial (research) team then this event might never be reviewed or reported in the trial or to the sponsor. It can also be within the publication itself where adverse events are poorly described or documented in the article itself. It has been suggested that this situation might be improved by the use of PROMs, especially those collected electronically to negate healthcare bias of interpreting patients' symptoms (92). Also there are ongoing efforts to encourage trialists to use CONSORT guidance to report the adverse effects more effectively in publication (89).

There is then natural concern that SAEs could be missed in the trial and hence potentially under reported (92). Separately, trials often do not continue long-term follow-up of patients due to the significant costs and administrative burden required to do this for many years. As such there is a concern that long-term SAEs which may have been caused by the drug are also missed; when the trial closes, trials also stop collecting SAEs and therefore an adverse condition that takes many years to evolve due to the drug may not be recorded (92). This was seen with zoledronic acid which can cause osteonecrosis of the jaw, but this association only discovered 2-5 years after FDA approval (93). Following approval of a drug the post licensing reporting of adverse events is at present suboptimal with under reporting by physicians using systems like the MHRA yellow card scheme within the UK. This makes post marketing review of SAEs difficult.

Large late phase cancer trials have the competing interest of gaining information of morbidity and mortality within the trial as they are often used for drug approval which trials such as phase 1 have less ability to review. Therefore, long-term side effects are

important particularly serious morbid events such as cardiovascular disease. Cancer therapies have a particular history of causing long-term cardiovascular effects be it anthracycline chemotherapy or radiotherapy to the chest (94). There is an ongoing concern that cardiovascular events in the short and long-term are being under reported in cancer clinical trials of FDA approved drugs. A recent study reviewed the cardiovascular outcomes of 189 late phase trials supporting cancer drugs from 1998 to 2018. They noted that over a third of the trials (37.6%) did not document any cardiovascular events. When cardiovascular outcomes were compared with non-cancer trial data for similar cohorts of patients there was a relative risk (RR) of a cardiovascular event of just 0.38 ($p<0.01$) in the trial cohort (95). Some have commented that this may just reflect the cohort of patients who were recruited (i.e. with minimal cardiovascular risk factors), however, this explanation seems unlikely due to the prevalence of cardiovascular disease in the relevant age group with patients with cancer. Therefore, understanding additional approaches for capturing these significant events within cancer trials is a priority.

2.1.2 EHRs and adverse event reporting

EHR, using centralised national databases, provide an additional resource that might be used to help understand the true number of adverse events in oncology clinical trials. Centralised national databases might have limited relevance in collection of 'non-serious adverse events' as most of these adverse events would not leave a record within these databases as they do not result in hospital admission. However, in the reporting and collection of SAEs that are life threatening and requiring hospital admission, it is possible that routine national centralised datasets could enhance trial data.

EHR data might be particularly helpful is in the long-term review of SAE data during the trial and following the closure of the trial. It could negate the problem of patient movement across the country as it has the potential to receive data from every hospital within the UK. The natural loss to follow-up that occurs within clinical trials could also be avoided. There have been examples of the use of routinely collected data in the past, however, it remains uncommon within UK trials to use this data (33).

It should be noted that EHR using centralised registry data is unlikely to replace *enhanced* reporting of SAEs during a trial as this needs to be reported within 24 hours to the sponsor. Centralised Datasets collect information from many different sources and then clean/ verify the data before it is published or distributed to third parties. This often means the lag time between the event and receipt of data by a researcher is often months to over a year after the event occurred. There is work to shorten this time frame but at present due to the significant time lag in the registry data this tight current regulatory deadline could not be feasibly achieved. For long-term follow-up however, routinely collected data could be a cheaper and more effective way to do this as it would potentially need minimal administrative staff at the sites or clinical trials unit (CTU) to manage.

There remains the important point that these centralised data bases were not designed for the use in clinical trials and the veracity of registry data needs to be reviewed to demonstrate to trialists, clinicians and regulatory authorities that data is equivalent to or even better then trial data before they can used and incorporated into trials consistently.

2.1.3 Cardiovascular electronic health data comparisons

EHRs using centralised registries have been used in cohort studies (as opposed to clinical trials) successfully in the past and continue to be used to good effect in population studies. This is an important distinction as the cohort may be defined as the group of patients for whom EHR is available (as opposed to clinical trials where patients have been specifically enrolled). The validity of this data applied at an individual (as opposed to population) level has been questioned however, which is of particular importance in relation to RCTs. A number of cohort studies have attempted to compare individual event data with that held within EHR to determine the strength of that data particularly in relation to cardiovascular outcomes. Cardiovascular events data has been thought to be particularly amenable to EHR collection as due to the seriousness of the condition nearly all the significant events need hospital admission to be managed and therefore will register within the EHR record.

The British Whitehall II study compared serial biomedical evaluations of cardiovascular events against HES and mortality data in a cohort of 7860 patients between 1997 and 2013 (96). In their study they ascertained information from self-reported events that was verified by Electrocardiogram (ECG) and General Practitioner and hospital events documentation. They compared their data on non-fatal coronary heart disease and stroke versus HES data from inpatient admissions using specified international classification of diseases (ICD-10) codes and Operating Procedure Codes Supplement (OPCS) coding. There was a total of 950 coronary heart disease events and 107 strokes. They found that there was reasonable concordance between the two sources with a sensitivity of 70% for coronary heart disease and 64-75% in stroke. They considered British Whitehall study documentation as the gold standard test for their sensitivity calculation. However, approximately 30% of the self/GP reported Whitehall data was not contained within HES. The authors concluded that this was due to poor quality of registry data in the earlier years of the study and angina events not requiring hospital admission.

Moving to data from a clinical trial context (where the tolerance of missing data at an individual patient level is much less), a systemic review of the comparison between cardiovascular trial outcomes and treatment effects used clinical endpoint committee adjudication versus routine data was presented at the Canadian 71st annual meeting of the Canadian cardiovascular society (only available to date in abstract form published with the Canadian journal of cardiology) (97). The data presented is a review of 7 comparative international studies comparing different routine data sets and outcomes. Due to the brevity of the abstract this review cannot be discussed in detail. However, it suggested in some of the studies a high level of agreement between the trial outcomes and routine data. From the minimal data that can be assessed there seems to be a wide range of publications involving different countries at different time periods. This is a severe limiting factor as comparison in this review may be limited as different routine data sources in different countries at different time points have a wide variety of effective linkage and data quality and as such the results may not be generalisable. If EHR data is to be used then the source that is appropriate for your study must be evaluated as registries have high degree of variability of linkage and reporting of certain outcomes.

The WOSCOPS data (discussed in the introductory chapter) is also pertinent here (61). Investigators compared their end point review committee trial data with routine health data in Scotland (98). This reviewed the concordance of routine collected data within Scottish morbidity records and also hospital admissions data with the original results of the WOSCOP study between 1988 and 1995. This comparison showed excellent matching between fatal cardiovascular events (97%) and over 80% for non-fatal myocardial infarction. When stroke and transient ischaemic attack (TIA) events were reviewed, concordance was slightly less at 78%. This paper also compared the outcome analysis using the two sets of data. This showed that even though there was difference in outcome for coronary heart disease (CHD) death and myocardial infarction (MI) this would have made no significant difference with respect to the overall qualitative outcomes of the study. However interestingly looking at CHD death, the risk reduction for CHD death with statin treatment would be slightly reduced, going from being marginally significant using the original or hospitalised WOSCOPS events (p=0.042) to marginally non-significant using the routine collected data events (p=0.093). In the data presented thus far, this is not reviewed in detail but could be potentially important for the impact of trials if an outcome goes from marginally significant to not, based on which data source is used. This is a trial based on Scottish registry data from some time ago and may not reflect today's EHR data particularly if a registry outside of Scotland is reviewed. In this evaluation they also concentrated on only events that needed hospital admission and excluded events that happened outside of hospital. This therefore also changes their outcome if they were to use registry data alone and not all the trial data when calculating their outcome.

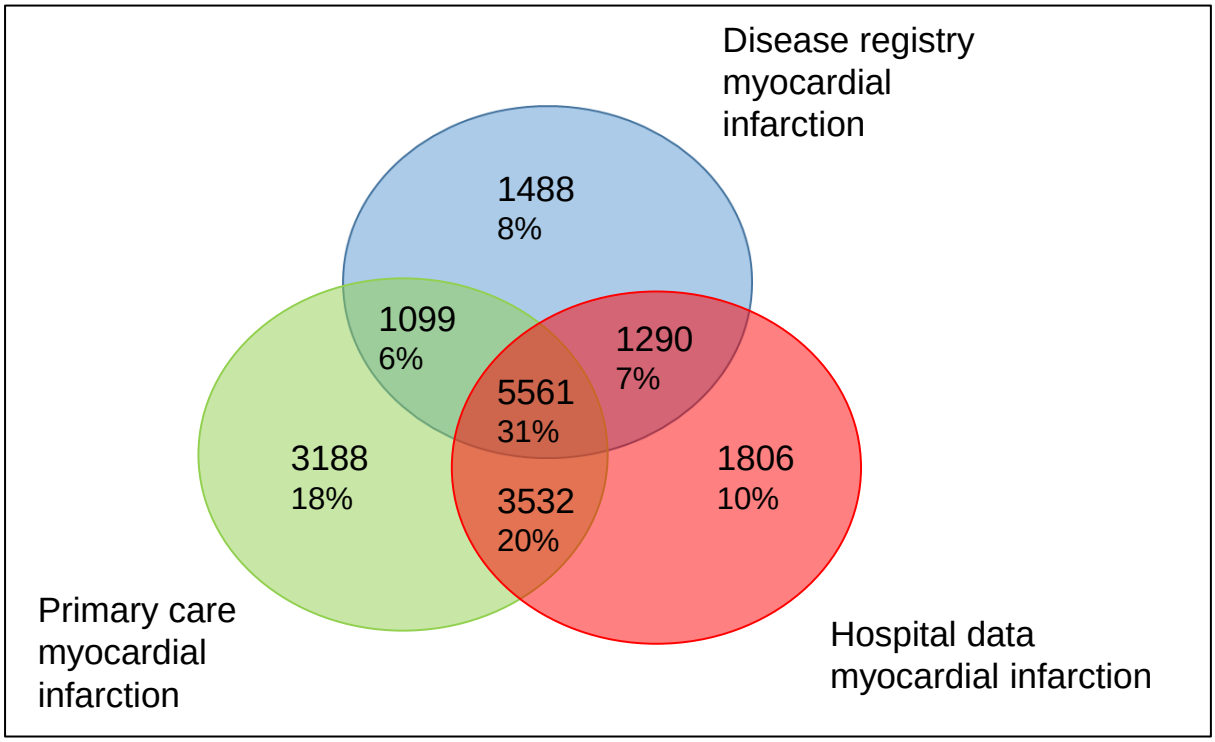
Perhaps the most comprehensive approach to date is that published in 2013 in the BMJ by Herret *et al*., which entailed a study based on three English national registries (79). The authors compared routinely collected GP data (from CPRD)) with HES (hospital admissions), the disease registry MINAP, and the ONS mortality register (cause specific mortality data). They compared data from all patients diagnosed with an acute MI from 2003 to 2009. The CPRD (discussed in the thesis introduction) covers EHR within a certain number of consenting GP practices. They established a cohort of patients who had consented to linkage of data that represented 3.9% of the population in England in 2006. They reviewed read codes with CPRD, ICD-10 codes

within HES and admissions for ST elevation and non-ST elevation MI within MINAP as defined by the internationally agreed definition of MI.

Events were compared between sources where there were no more than 30 days different in reporting cardiovascular event occurrence. They assessed agreement between the three sources and presented the data using a Venn diagram for patients suffering acute MIs who survived more than 7 days. They also calculated the positive predictive value (PPV) of primary care or hospital discharge diagnoses of acute MI among patients who also had a record in the acute coronary syndrome (ACS) registry. To understand if there were demographic reasons for rate of reporting they also performed a logistic regression analysis to establish if age, sex, deprivation, rate of primary care consultation, year of myocardial infarction, or mortality at 30 days explained suboptimal recording of acute MI in primary care, hospital discharge, or disease registry sources. An additional analysis compared patients who died within 7 days with the knowledge that a full comparison could not be made as hospital admission data would not have registered patients that died prior to arriving to hospital.

21,482 patients who had fatal and non-fatal MI were included. The authors demonstrated that in 20,000 patients each data course missed a substantial proportion of myocardial events. Of non-fatal myocardial events, only a third were recorded in all three data sets. The CPRD was the single most complete source of non-fatal MI records (where one quarter of all non-fatal myocardial infarction events not recorded), HES missed one third, and MINAP missed nearly half (**Figure 2.1)**.

***Figure 2-1: Venn diagram of comparison of reporting of non-fatal myocardial infarction in 17964 patients 2003-2009 between different EHR data sources (community, hospital and disease registry) adapted for the purposes of this thesis from Herret BMJ 2013 publication* (79)**



This study demonstrated the potential importance of using linked data incorporating multiple sources of information to get true picture of total rate of cardiovascular disease within a population.

2.1.4 PATCH trial

The focus of the work detailed within this chapter, compares cardiovascular events collected within the Prostate Adenocarcinoma TransCutaneous Hormone trial (PATCH, MRC PR09 (ISRCTN70406718)) trial with a national English registry data held by NHS digital (HES) and the NICOR national audits.

PATCH is a randomised trial for men with locally advanced or metastatic prostate cancer designed with a staged approach to evaluate tE2 in this setting. The primary aim is to test whether tE2 are non-inferior to LHRHa in terms of overall survival and progression free survival. Importantly, tE2 may have an improved toxicity profile which is key as patients with prostate cancer can live for many years on treatment (99). See trial schema below (**Figure 2.2)**

The backbone to prostate cancer treatment is the depletion of testosterone as testosterone is required for cellular growth of prostate cancer (100). Prostate cancer was initially treated with oral oestrogens to suppress testosterone from the 1950s to 1980s, demonstrating good control of the cancer (101). This was achieved through the administration of exogenous oestrogen that, through a negative feedback loop on the hypothalamus and pituitary, achieve castrate levels of testosterone and avoids the physiological effects of oestrogen depletion (**Figure 2.3)** (102, 103). However orally administered oestrogens were associated with significant cardiovascular and pro-thrombotic effects and were superseded by LHRH analogues (104, 105). LHRHa also suppressed the testosterone with equivalent treatment effects on prostate cancer. However as with many treatments there are now increasing concern of the long-term side effects. LHRHa increases the risk of metabolic syndrome and osteoporosis (**Figure 2.3)** (106, 107). This may be in part because LHRHa also stops the production of oestrogen which contributes to these two long-term side effects. These can both cause life threatening conditions with fractures or potentially the increased risk of diabetes and also cardiovascular disease (108-110).

The PATCH study uses tE2 instead of oral oestrogen. The cardiovascular and pro-thrombotic complications of oral oestrogens are thought to be induced by first pass effect in liver metabolism which tE2 avoids (111). Therefore, the risk of cardiovascular disease is thought to be considerably reduced. It is proposed that tE2 would provide better initial and long-term side effect profile compared to LHRHa and also achieve comparable effect on the treatment of prostate cancer. This has been demonstrated with early data from the PATCH trial which reviewed quality of life between transdermal oestradiol and LHRHa. This demonstrated better self-reported quality of life particularly with respect to hot flushes and fatigue (99). The trial has also demonstrated better bone health and also metabolic outcomes for transdermal oestradiol versus LHRHa (112, 113).

The PATCH trial from its concept has had one of its major trial outcomes as cardiovascular events. As this was crucial to the long-term confidence in repurposing tE2 for the treatment of prostate cancer, PATCH has employed enhanced clinician review of cardiovascular SAEs. This meant that all necessary evidence of these events, through investigation reports and clinical notes, were asked for and reviewed against a set end-point definition of a cardiovascular event. This was designed to ensure that cardiovascular events that were reported were true events and not just ‘possible’ events with no definitive evidence. Another key to the comprehensive review of cardiovascular outcomes within PATCH was that even if the participant was not on the study drug (e.g. having switched to LHRHa due to side effects or cancer progression) the protocol required investigators to report all potential cardiovascular events until closure of the trial. This is different to many other trials where these events were potentially only reported for 30 days following the cessation of the investigated medicinal product.

Figure 2-2: PATCH trial schema directly copied from PATCH protocol Version 13 (March 2020) (99)

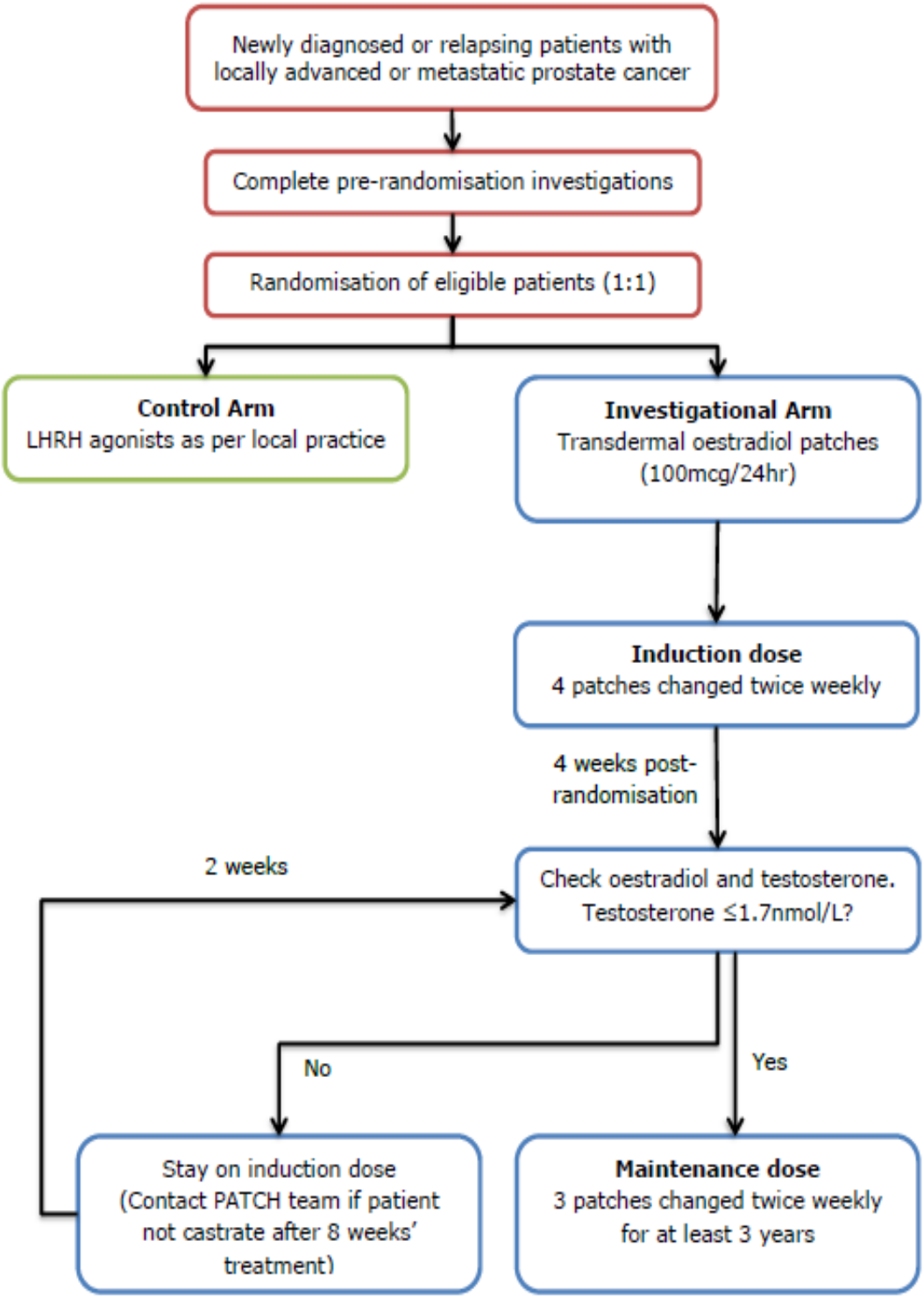
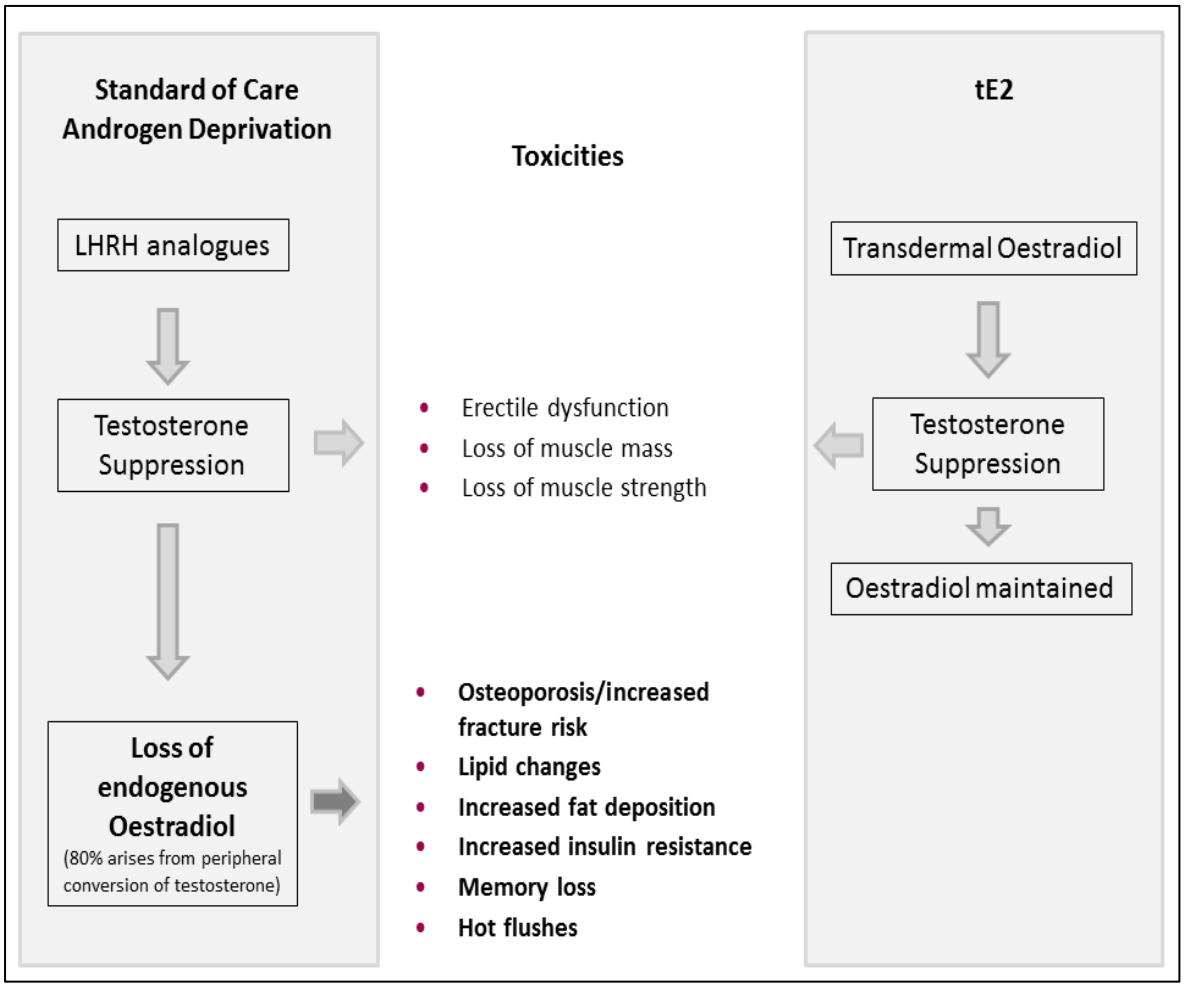


Figure 2-3: Toxicities of androgen suppression with LHRHa attributable to low testosterone vs suppression of endogenous oestrogen



These two factors (enhanced reporting and continued review of cardiovascular events) mean PATCH is a unique opportunity to compare these events to EHRs. Notably, with respect to comparing cardiovascular events between tE2 and standard of care LHRHa, in an early analysis of 254 patients the PATCH team demonstrated through this enhanced reporting that there was no significant difference of cardiovascular events between LHRHa and oestradiol patches (114). This allowed the trial to develop into a phase 3 trial to rigorously review the longer term effect of cardiovascular events and also see that oestradiol patches were not inferior to LHRHa in the treatment of prostate cancer in the locally advanced and metastatic setting. The results of the long-term follow-up of cardiovascular events have now been published demonstrating no difference to first cardiovascular event between the two treatments in 1694 men (HR 1.11 CI 0.80-1.53 p=0.54) (115).

Therefore, cardiovascular outcomes within PATCH provide an excellent substrate to replicate and update the approach taken by Herrett *et al*., where they set out to compare three sources of data. The unique difference in this study is that one routine source of data is derived from clinical trial data (from PATCH). This required accessing data from NHS digital (HES) and also NICOR cardiovascular audit data. These two holders of data (as described in the thesis introduction) were considered at the time the best routine data for patients admitted to hospital with cardiovascular event. These are both potential sources of data that could be used to replace or enhance long-term trial data. In this chapter we will describe the comparability of this data but also the feasibility of this data being used as a replacement for SAEs in trial data reviewing lag time of routine data reporting and accessibility. This would allow clinical trialists in the future to consider if both data sources are needed to collect long-term data or just one alone or neither to collect trial outcomes data.

2.2 Methodology

2.2.1 Study Outcomes

The aims of this study were as follows:

2.2.1.1 Primary Outcome

1. Analyse the concordance of all cardiovascular events between NHS Digital HES APC data and enhanced collection of cardiovascular events within the PATCH trial
2. Analyse the concordance of acute coronary syndrome and heart failure events between NICOR MINAP and Heart Failure audits, NHS Digital HES APC data and enhanced collection of cardiovascular events within the PATCH trial

2.2.1.2 Secondary Outcome

1. Analyse the concordance of all cardiovascular events between NHS Digital HES A&E data and enhanced collection of cardiovascular events within the PATCH trial

2.2.2 Ethical Approval

The PATCH trial was approved in Nov 2005 by the Leeds (East) Research Ethics committee. The PATCH methodological sub-study investigating EHR was ethically approved within protocol version 11 in Feb 2019. The project design and timing was also agreed by the IDMC /Trial Steering Committee and trial funders. Patient information sheets and consent forms were felt to be sufficient by NHS Digital and NICOR, however, consent to nationally held data could only be confirmed for participants after 2010 when this question became mandatory rather than optional. Therefore, participants recruited after 2010 were only included in this methodology study. A transparency/ privacy notice was also written to participants as per data

protection law to describe exactly how their routine data would be used within the trial. This was sent out to each site to update their participants and made publicly available on the PATCH trial website. (**Appendix A**)

2.2.3 Dataset Definitions

The three datasets involved in the comparison were the PATCH trial dataset of cardiovascular events according to protocol definitions, HES data from NHS Digital and MINAP/Heart failure national audit data from NICOR. For all three datasets, definitions and rules were applied for appropriate analysis.

2.2.3.1 PATCH trial dataset

Cardiovascular (CVS) events were defined as a specific end-point of interest in the PATCH protocol. Serious and notable events that were considered potentially CVS events, by either the site investigator or clinical reviewer, underwent additional review by the PATCH team. All available evidence was reviewed by the clinical reviewer to determine if the event met the criteria to be considered a CVS outcome event, as per protocol definitions. The outcome of this review was subsequently recorded on a specific CVS end-point review form.

In PATCH the cardiovascular endpoint definitions that were defined in the protocol are as follows:

1. **Heart Failure:** new symptoms or clinical signs consistent with a diagnosis of new or decompensated cardiac failure with supporting evidence from Chest X-Ray, ECHO or rise in BNP.
2. **Acute Coronary Syndrome (ACS) (including unstable angina, NSTEMI, MI):** new onset cardiac chest pain, confirmed as ischaemic in origin by ECG and/or troponin rise +/- coronary angiography. In the case of collapse or new shortness of breath associated with a silent infarct, the latter would also be confirmed by ECG and/or troponin rise +/- coronary angiography.

3. **Thromboembolic stroke:** new neurological symptoms and signs consistent with a CVA with confirmatory evidence from brain CT or MRI. For transient ischaemic attacks (TIA), the diagnosis will be clinical, with corroborative data from carotid duplex scanning. Evidence will also be sought for pre-existing or new, persistent or paroxysmal, atrial fibrillation.
4. **Other arterial embolic events:** detected by new clinical symptoms and supporting radiological evidence.
5. **Venous thromboembolism**: Thromboses confirmed radiologically or pulmonary embolism (PE) confirmed by means of CT pulmonary angiogram (CTPA). In rare cases, depending on clinical circumstances, the confirmation may be by ventilation/perfusion scans or angiography.
6. **Death attributed to any of the above** (where the event was not documented according to the definitions provided above).

Prior to analysis all CVS events and CVS "non-events" underwent a final clinical review, by Prof Ruth Langley, Dr Duncan Gilbert and Dr Archie Macnair, to ensure consistency of reporting, bearing in mind the CVS endpoint review had been carried out over many years by multiple clinical reviewers.

This re-review process highlighted that there were minor inconsistencies in reporting suspected TIA and acute coronary syndrome as CVS events. The re-review took a conservative approach and included any events that clinically matched the definitions above or had sufficient evidence to change from a CVS "non-event" to a CVS event.

Therefore, trial clinicians at the CTU reviewed all CVS "non-events" under these categorises and changed the following:

1. 5 TIAs changed from CVS "non-event" to CVS event
2. 2 ACS changed from CVS "non-event" to CVS event
3. 1 Heart failure episode changed from CVS "non-event" to CVS event

All CVS events that resulted in death were reviewed. Any death with evidence confirming a CVS event was defined as per the appropriate category 1-5. Deaths with a cause that did not meet the criteria were defined as CVS "non-event".

Any sudden unexplained death of a patient where evidence from a postmortem supported a CVS cause of death, was assigned to the appropriate CVS event category. All sudden unexplained deaths where there was no postmortem were re-reviewed prior to publication by trial clinicians at the CTU. Following this review 10 events were still deemed to have no clear other cause of death and died outside of hospital. These events were categorised as CVS "non-event" and were excluded from this analysis. Previously these 10 events were included in a sensitivity analysis of cardiovascular outcomes but were excluded from this analysis as the death was not definitely a CVS event and as occurred outside of hospital and would not register within HES or NICOR.

2.2.3.2 NHS Digital HES Data

Potential areas of data within HES are those derived from A&E, APC, outpatient and critical care. We felt that the most appropriate comparison would only be with HES APC. HES critical care and outpatient data should not be expected to add any additional significant information in this setting. For the primary analysis PATCH trial data endpoints defined as Heart failure, ACS, cerebrovascular or other arterial thromboembolic events or venous thromboembolic disease were compared with HES APC data alone in a two way comparison. For the primary analysis comparing the PATCH endpoints defined as heart failure and ACS with HES APC and NICOR which was a three way comparison between the defined datasets. Comparison would be made with HES APC diagnosis box fields which hold diagnosis for the patients' admission using ICD-10 codes. There are 20 diagnosis boxes. Diagnosis box 1 is the primary diagnosis for that admission to hospital with the other fields containing secondary (contributing acute reasons for admission) or subsidiary diagnosis (containing previous or important chronic conditions relevant to the hospital admission).

HES A&E data was also considered for this comparison as many thromboembolic events – especially deep vein thromboses (DVT) and minor thromboembolic strokes would be managed without admission to hospital but just by A&E attendance. There

were limitations to this analysis that some of these events will be managed entirely as an outpatient so will not be picked up by the A&E dataset. Outpatient data was not included as the detail in the dataset would be not sufficient to establish if they were treated for either of these conditions during an outpatient consultation. The second limitation was A&E data may have been subject to poorer diagnosis coding and also use alternative codes to HES APC data, using A&E coding instead of ICD-10 codes (116, 117). As this was the only way that these cardiovascular events could be picked up in the HES data it was considered appropriate to do a separate analysis with A&E data and PATCH trial data to see if there was a degree of comparability between the two or if events could be picked up in A&E data that were not picked up in trial data. This would also assess the degree of reliability of A&E data coding to see how it could be used for trials in the future. A&E data would not be compared with the HES APC analysis in the primary analysis.

2.2.3.3 NICOR dataset

NICOR holds six audits: Adult cardiac surgery; Adult coronary percutaneous intervention; Cardiac rhythm management; Congenital heart disease in children and adults; Heart failure; and MINAP audit (47). Description of NICOR datasets history and have been described in the thesis introduction. Data supplied by the audits differed from HES data as it had much greater detail on the clinical event. This supplied symptoms/ risk factors, investigation details (blood test/ ECG)/ echocardiogram results, treatment (medication or procedure), timing of procedures and end result of admission.

Only the PATCH end point definitions of heart failure and ACS met the criteria for comparison with the NICOR audits. The most relevant of those audits were considered to be MINAP and Heart failure. Although the Adult cardiac surgery and Adult coronary percutaneous intervention may have given extra information, they pertained to elective intervention (and included e.g. repair of valvular heart disease that is not relevant to questions around tE2) and as such were not included. Other endpoint definitions were not compared within the NICOR audits.

The heart failure audit event inclusion is defined as all those patients with an unscheduled admission to hospital in England and Wales who are discharged with a

primary diagnosis of heart failure. This is also defined on discharge of a primary diagnosis of heart failure based on the ICD-10 codes: (118)

I110 Hypertensive heart disease with (congestive) heart failure
I255 Ischaemic cardiomyopathy
I420 Dilated cardiomyopathy
I429 Cardiomyopathy, unspecified
I500 Congestive heart failure
I501 Left ventricular failure
I509 Heart failure, unspecified

The MINAP audit is a national audit of all admissions admitted to hospital with acute coronary syndrome. Over the last two summary reporting years (2018/2019) the MINAP audit has been including in the annual report data verification between HES data and data submitted to the audit to demonstrate hospital variation in reporting based on defined ICD-10 discharge codes. In the last annual report of 2019 at time of writing (years covering 2017-2018) they have defined inclusion into the audit using the following ICD-10 codes as: (119)

STEMI: all patients discharged with final diagnosis of STEMI – identified by the presence of the following ICD 10 codes in ANY position:

I21.0 ST elevation (STEMI) myocardial infarction of anterior wall;
I21.1 ST elevation (STEMI) myocardial infarction of inferior wall;
I21.2 ST elevation (STEMI) myocardial infarction of other sites;
I21.3 ST elevation (STEMI) myocardial infarction of unspecified site.

NSTEMI: all patients discharged with final diagnosis of NSTEMI – identified by the presence of the following code in the FIRST position:

I21.4 Acute subendocardial myocardial infarction.

MINAP would only use events for their annual report that met this criteria on ICD-10 codes. This would only affect a minority of the cases within this comparison and therefore did not affect the decision on which codes to be used in HES APC in this

comparison as previously they used with a broader definition of ICD-10 coding or based on a clinician definition.

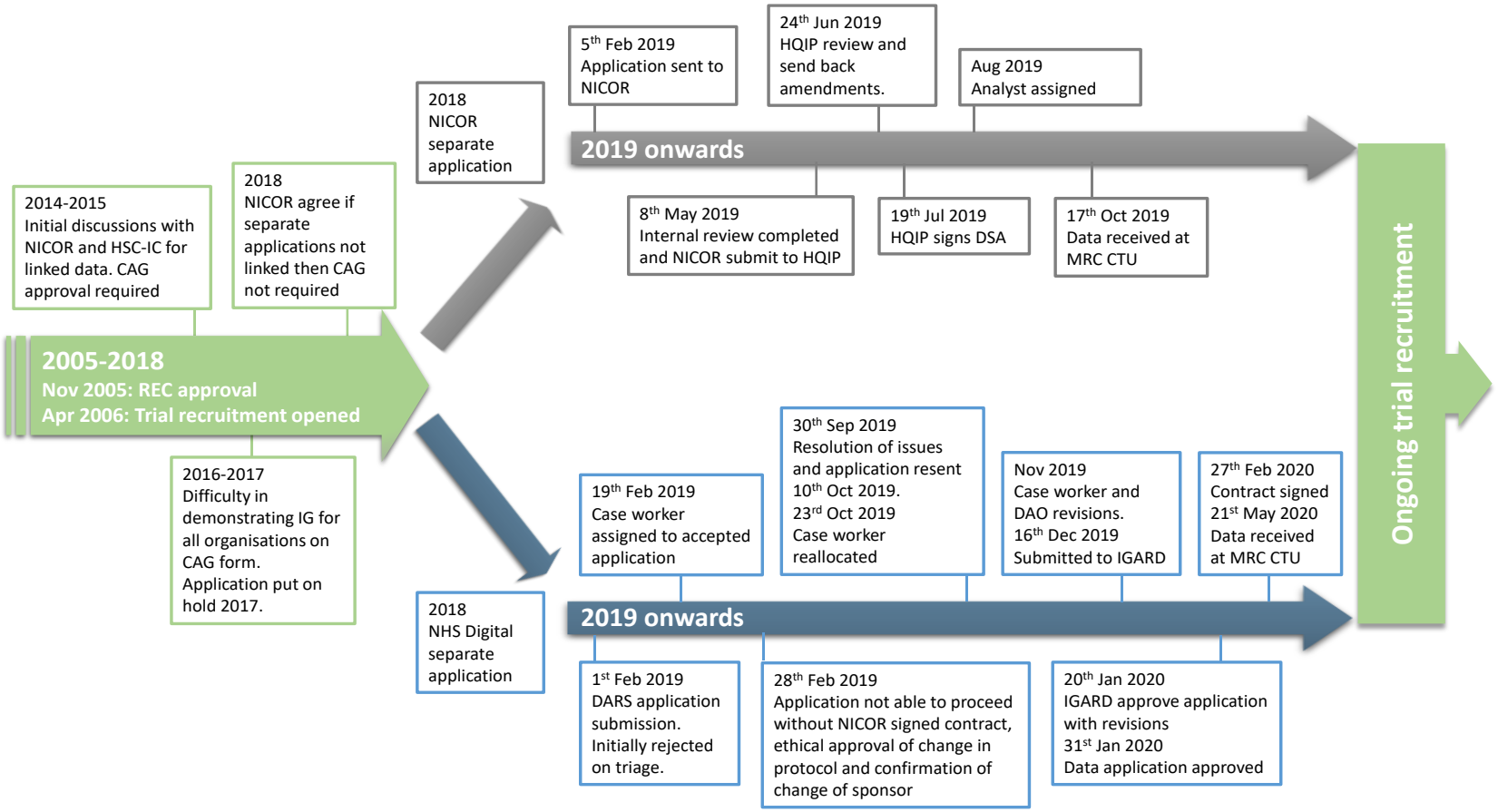
2.2.4 Data access

As a general principle, access to the national data registries is based on the scope of the project, legal basis, degree of anonymization of data and trial documentation. The narrative of how data was obtained from the two datasets has been discussed with reflections and is in **Appendix B** and also published (120). A summary of the timeline is stated in **Figure 2.4**.

Historic details around the wording of consent taken at participant randomization conferred some limitations on the breadth of data sought. Initially it was proposed that NICOR and NHS Digital data would be linked by NHS Digital to give the best chance of all events being received and analysed by the MRC CTU. However, participant's consent at the time was thought not to be sufficient for linked data between NHS Digital and NICOR. A Confidentiality Advisory Group (CAG) approval was a prerequisite to allow linkage of the two datasets by NHS Digital but this was not possible due to administrative reasons within PATCH and timescales of the project. Therefore, separate applications were submitted to both with the triangulation project proposal but not to share NICOR data directly with NHS Digital.

Participants prior to 2010 were also not included in this analysis as it was not possible to confirm the consent for the use of participants' registry data in further research. Prior to 2010 this question was not a mandatory field for completion for subjects to enter the trial. Consent forms were not held by the CTU so collection of these approximate 250 consent forms (prior to 2010) from individual sites was not considered feasible for this research project. The analysis would therefore be a subpopulation of the PATCH cohort starting from 2010, and including only English participants in the trial (as NHS Digital only covers England).

Figure 2-4: Flow diagram of the PATCH joint application to NHS Digital and NICOR and subsequently handled as separate applications in 2018 (Please note that timeline is not proportional). Adapted for the purposes of this thesis from Macnair Trials 2021 publication. (120)



CAG: Confidentiality Advisory Group; DAO: Data Approvals Officer; DARS: Data Access Request Service; HQIP: Health Quality Improvement Partnership; HSC-IC: Health and Social Care Information Centre; IGARD: Independent Group Advising on the Release of Data; REC: Research ethics committee

2.2.5 Censorship dates

Importantly, data providers incorporate variable lag times with respect to available data. For HES data this includes data up to the nearest month following a data sharing agreement April 2020. For NICOR only published data can be released. As such the MINAP audit had a censorship date of January 2018 with the heart failure audit being censored at 31 March 2018.

Therefore, the analysis encompasses cardiovascular events that started in 2010 following recruitment of patients with valid consent into the trial to January 2018 so that a fair comparison was completed between all three. Despite the different times in collecting data, as this was up to January 2018 it was deemed that all datasets were comparable in their data freeze.

2.2.6 Data Processing/ Information governance

Both data providers were sent identifiable data for a cohort of participants as stated above that they could send their data on the relevant participant from the datasets. This identifiable data was held in a separate database to the PATCH trial database and managed Mary Rauchenberger, the head of data management systems, who facilitated the transfer of data.
The identifiable information was:

- Study ID
- NHS Number
- Name
- Date Of Birth
- Gender
- Postcode

The data flow for the project is shown in **Figure 2.5**:

Figure 2-5: Dataflow diagram from NHS Digital and NICOR to MRC CTU at UCL

All data was transferred as per the data requirements set out in the data agreements for both datasets and held within the UCL data safe haven. No identifiable data was held within the project specific section of the data safe haven and all data from NHS Digital and NICOR was pseudo-anonymised and no participants data was be re-identified in the analysis or for publication. No NICOR audit or NHS Digital data was removed from the data safe haven. Data was not integrated between the three datasets but separate comparisons were made between each to complete the analysis. Only defined members of UCL staff who work on the PATCH study or MRC CTU methodology group were allowed access to the data if they had proof of information governance training and data safe haven training. Data was not shared with a third party and data from the sub-study would not be used to inform the outcomes of the study.

There was a slight difference in patient inclusion between the two datasets in that if they had had a hospital/ A&E admission then we would receive all information for those participants in the HES data no matter what the cause of the admission. This was to make sure that all hospital admission events that did happen could be assessed for the methodology project. NICOR MINAP and Heart failure audits only hold information on participants who have had either an acute coronary event or heart failure event.

2.2.7 Analyses

For each comparison, datasets and events were defined as above but specifically to each analysis:

2.2.7.1 Primary outcome analysis all CVS events

1. Within the trial data every event that had been classed into the categories of a cardiovascular endpoint were included apart from sudden unexpected death. All CVS ‘non-events’ were excluded. In this review HES events were compared with all PATCH trial endpoint events including heart failure, ACS,

cerebrovascular or other arterial thromboembolic events or venous thromboembolic disease.

2. HES APC was compared with all cardiovascular outcomes apart from trial coded in the category 'death'. This was initially defined as ICD-10 codes that populate the first diagnosis box only to see the difference in PPV and sensitivity between the two methods. There was a subsequent analysis using the first 5 diagnosis boxes. The ICD-10 codes for each event are as follows in **Table 2.1**.

2.2.7.2 Primary outcome analysis heart failure and ACS only

1. Within the PATCH trial data, every event that had been classed into the categories of a cardiovascular endpoint were included apart from sudden unexpected death. All CVS 'non-events' were not included. In this review the NICOR and HES events were only compared with acute coronary syndrome and heart failure.
2. Within NICOR all events that were received from the MINAP and Heart failure were included as long as they were within the confirmed trial time period for each participant. NICOR does not use coding for this and data fields are explained and populated as per their data dictionary.
3. HES events that included acute coronary syndrome and heart failure events in the HES APC data were included in this analysis. This was initially defined as ICD-10 codes that populate the first diagnosis box only. There was a subsequent analysis for diagnosis box 1-5 to understand the difference in PPV and sensitivity between the two methods. The codes were defined on previous published validation work and also NICOR definitions. (79, 118, 119) The ICD-10 cardiovascular codes are shown in **Table 2.2**.

Table 2-1: ICD-10 codes for two way comparison

PATCH CVS Diagnosis	HES: ICD-10 code version 2019 and corresponding diagnosis with ICD-10 code
Acute coronary syndrome	I21* Acute myocardial infarction I22* Subsequent myocardial infarction I23* Current complications following acute myocardial infarction I249 Acute ischaemic heart disease, unspecified
Heart failure	I110 Hypertensive heart disease with (congestive) heart failure I255 Ischaemic cardiomyopathy I420 Dilated cardiomyopathy I429 Cardiomyopathy, unspecified I500 Congestive heart failure I501 Left ventricular failure I509 Heart failure, unspecified
Thromboembolic stroke	I63* Cerebral Infarction I64* Stroke not specified as haemorrhage or infarction (thromboembolic stroke) G45* Transient cerebral Ischaemic attacks and related syndromes
Venous thromboembolism	I26* Pulmonary Embolism I802 Phlebitis and thrombophlebitis of other deep vessels of lower extremities incl deep vein thrombosis NOS 1803 Phlebitis and thrombophlebitis of lower extremities, unspecified including embolism or thrombosis of lower extremity NOS I81* Portal vein thrombosis I82* Other venous embolism and thrombosis
Other arterial embolic event	I74* Arterial embolism and thrombosis

**indicates that all four digit codes used with starting three digits*

Table 2-2: ICD10 codes for triangulation analysis

PATCH CVS Diagnosis	HES: ICD-10 code version 2019 and corresponding diagnosis with ICD-10 code
Acute coronary syndrome	I21* Acute myocardial infarction I22* Subsequent myocardial infarction I23* Current complications following acute myocardial infarction I249 Acute ischaemic heart disease, unspecified
Heart failure	I110 Hypertensive heart disease with (congestive) heart failure I255 Ischaemic cardiomyopathy I420 Dilated cardiomyopathy I429 Cardiomyopathy, unspecified I500 Congestive heart failure I501 Left ventricular failure I509 Heart failure, unspecified

**indicates that all four digit codes used with starting three digits*

2.2.7.3 Secondary outcome analysis

HES A&E data was compared with all cardiovascular outcomes apart from death. This was defined by using the first diagnosis box and either using ICD-10 codes as per the HES APC analysis or HES A&E coding. Only events with valid A&E coding were used for a comparison analysis.

A&E coding for diagnosis uses a six-character code with diagnosis condition (2n) sub analysis (1n) anatomical area (2n) and then anatomical side (1n). HES A&E events will be matched with trial cardiovascular event points using the following codes at the start of valid codes as per A&E data dictionary (121):

- 20; cardiac condition
- 21; cerebrovascular condition
- 22; vascular condition
- 201; myocardial infarction
- 202; other cardiac condition would be excluded

Examples of coding includes:

20122L- 201 (Cardiac conditions - myocardial ischaemia & infarction) 22 (Chest) L (Left) - Myocardial infarction

22832R- 22 (Other vascular condition) 8 (filling character) 32 (leg) R - DVT right leg

2.2.8 Statistical Analyses

2.2.8.1 Primary outcome analysis all CVS events

This was a direct comparison between HES APC data and trial cardiovascular endpoints using the codes and definitions described above based on matching via STATA. This was based on a defined cardiovascular event within the trial data and HES APC (as per coding of event above) with exact matching date of admission or with a 2 week window either side. The events were censored so that events that only occurred within the study period or following the participants recruitment were analysed. The analysis also censored for multiple records for the same event. If there

were two significant events in the NHS data then both would be compared with trial data giving two separate data points.

Two separate analysis were performed firstly using HES data from diagnosis box 1 and then diagnosis box 1-5. This was carried out by Matthew Nankivell (NM), PATCH trial senior statistician. Tables and graphs were created for both with PPV/ sensitivity calculations with trial data defined as the gold standard/ reference standard carried out by Archie Macnair (AM). Trial data was used as the 'gold standard' as this data would have been used in the publication of the trial. This would describe the disparity between trial only event and no registry recorded event (sensitivity) and registry recorded event but no matching trial event (PPV). Decision to compare diagnosis box 1-5 versus diagnosis box 1 rather than diagnosis box 1-20 was based on preliminary analysis of cardiovascular events to demonstrate the potential difference in PPV and sensitivity in using the two difference comparisons with trial data. If all 20 boxes were used then the number of events recorded was dramatically different from just box 1 and unlikely to be the cause of the admission if the 20th reason. Therefore box 5 was used as near the median for frequency of diagnosis pick up.

HES APC data were also reviewed against cardiac events that did not meet trial cardiovascular definition and SAE data to see if there were confirmed matches. Following this AM reviewed each event that matched with a Cardiac 'non-event' and SAE data. This review described any common themes or explanations for this match. Lastly AM went through any event that did not link to any of the other datasets to understand any common themes or explanation for no matches.

2.2.8.2 Primary outcome analysis heart failure and ACS only

Venn diagrams were created by AM for the triangulation of the three datasets based on matching via STATA carried out by NM. This was based on a defined cardiovascular event within the trial data (as above) for acute coronary syndrome and heart failure, NICOR event and HES APC (as per coding of event above) with exact matching date of admission or with a 2 week window either side. The events were censored so that events that only occurred within the study period or following the participants recruitment were analysed. The analysis also censored for multiple

records for the same event. If there were two significant events in the NHS data then both would be compared with trial data giving two separate data points.

Two separate analysis were performed firstly using HES data from diagnosis box 1 and then diagnosis box 1-5. This was carried out by Matthew Nankivell (NM), PATCH trial senior statistician. Tables and Venn diagrams were created for both with PPV/ sensitivity calculations with trial data defined as the gold standard/ reference standard carried out by Archie Macnair (AM). Trial data was used as the 'gold standard' as this data would have been used in the publication of the trial. This would describe the disparity between trial only event and no registry recorded event (sensitivity) and registry recorded event but no matching trial event (PPV). Decision to compare diagnosis box 1-5 versus diagnosis box 1 rather than diagnosis box 1-20 was based on preliminary analysis of cardiovascular events to demonstrate the potential difference in PPV and sensitivity in using the two difference comparisons with trial data. If all 20 boxes were used then the number of events recorded was dramatically different from just box 1 and unlikely to be the cause of the admission if the 20th reason. Therefore box 5 was used as near the median for frequency of diagnosis pick up.

HES APC and NICOR data were also reviewed against cardiac events that did not meet trial cardiovascular definition and SAE data to see if there were confirmed matches. Following this AM reviewed each event that matched with a Cardiac 'non-event' and SAE data. This review described any common themes or explanations for this match. Lastly AM went through any event that did not link to any of the other datasets to understand any common themes or explanation for no matches.

2.2.9 Secondary Analysis

A separate analysis was carried out with cardiovascular events within the HES A&E dataset and trial data. This was not linked to the previous HES APC analysis. This was for all cardiovascular event point events apart from death using the HES A&E codes defined above. Previous data from NHS Digital and previous studies stated a high level of inaccuracy of HES A&E diagnosis coding (116, 117) so the initial analysis was to see how many diagnosis codes were missing and also how many diagnosis codes stated 'diagnosis non classifiable' defined as HES A&E codes 38. The Initial review

compared the two datasets with the stated HES A&E codes. If there was a large disparity then the exact matches were reviewed to see if there was any consistency of HES A&E coding with the known clinical trial events. This was in order to see if the HES A&E coding were uniform enough compared with trial events so that this could be used as a potential dataset for trials in the future.

2.3 Results

1,200 English patients on the PATCH study were reviewed between January 2010 to January 2018. There were 71 cardiovascular events, broken down into the separate categories, are seen in **Table 2.3**. This reflected 71 of 157 of the potential total cardiovascular events described within the PATCH trial data (2006-2019) at the time of analysis.

HES APC data was reviewed within the same period and 12,500 separate hospital events were analysed using the ICD-10 codes as above with the number of cardiovascular events listed by number of diagnosis box reviewed either diagnosis box 1 only or diagnosis box 1-5. This meant there were 57 events in total for HES diagnosis box 1 and 136 events in total with HES diagnosis box 1-5. NICOR data had 49 events within MINAP and 15 events from the Heart failure audit. Events that occurred prior to patient trial randomization were also excluded. Following the removal of these events the number of cardiovascular events available were 9 within MINAP and 15 with the heart failure audit.

The most significant differences observed when expanding the definition to include boxes 1-5 is the increase in number of heart failure events (from 14 to 65) although there were also a number of additional venous thromboembolic events included (from 11 to 27).

Initially all HES APC events were compared with trial cardiovascular events that met the pre-specified definitions. This overview comparison is summarised in **Table 2.4** and **Figures 2.6 and 2.7** for HES APC diagnosis box 1 comparison and HES APC diagnosis box 1-5 comparison.

Table 2-3: Cardiovascular events derived from patient data from the clinical trial database (CTD), HES APC dataset (HES) and NICOR for patients enrolled in PATCH between 2010-2018. NICOR audits include heart failure and ACS data only.

CVS Event	Number of events			
	CTD	HES diagnosis box 1	HES diagnosis box 1-5	NICOR
Heart Failure	15	14	65	15
Acute Coronary Syndrome (ACS)	18	15	22	9
Thromboembolic Stroke	17	15	18	N/A
Other arterial embolism	2	2	4	N/A
Venous Thromboembolism	19	11	27	N/A
Total	71	57	136	24

Table 2-4: Events on CTD and HES APC dataset (HES)

	Diagnosis box 1 alone			Diagnosis box 1-5		
	CTD alone	HES1 only	Both	CTD alone	HES1-5 only[4]	Both[4]
HF	11	10	4	9	59	6
ACS	10	7	8	8	12[2]	10
Stroke	9	7[1]	8	8	9[1]	9
Arterial embolism	2[3]	2	0	2	4	0
Venous thromboembolism	12	4	7	9	17	10
Total	44	30	27	36	101	35

1. One event is on a trial CRF, with diagnosis “Other arterial embolism”.
2. One event is on a trial CRF, diagnosis “Heart failure” (this event has both HF and ACS diagnoses within HES box 1-5).
3. One event is on HES box 1, diagnosis “Stroke”.
4. Five events cover two diagnosis in HES box 1-5, and are listed under both categories.
    a. Three are not on a trial CRF – 2 events are ACS (box 1) and HF (box 2-5); 1 event has no diagnosis (box 1) and HF and arterial embolism (box 2-5).
    b. 1 on a trial CRF – 1 is HF on CRF, and HF (box 1) and ACS (box 2-5)

2.3.1 Trial data/ HES comparison for all cardiovascular events

Overall 27/71 (38.0%) cardiovascular events from the clinical trial database were present in HES data using diagnosis box 1, rising to 35/71 (49.3%) if HES diagnoses 1-5 were used. However, there were a similar number of events again (n=30) described in the HES data (diagnosis box 1) that were not seen in the clinical trial database. If a more permissive definition (HES diagnosis 1-5) was used this rose to an additional 101 events. This was particularly noticeable for heart failure where this diagnosis appeared an additional 59 times in HES boxes 1-5, but all cardiovascular outcomes demonstrated numerous events that appeared in one or other of the datasets but not both.

Considering the clinical trial data as the gold standard, for events detected in HES diagnosis box 1, this translated to a positive predictive value (PPV) of 0.47 and a sensitivity of 0.38. Using HES diagnostic boxes 1-5 however this dropped to a PPV of 0.26 although with improved sensitivity at 0.49.

Figure 2-6: CTD data vs HES APC dataset (HES) diagnosis box 1

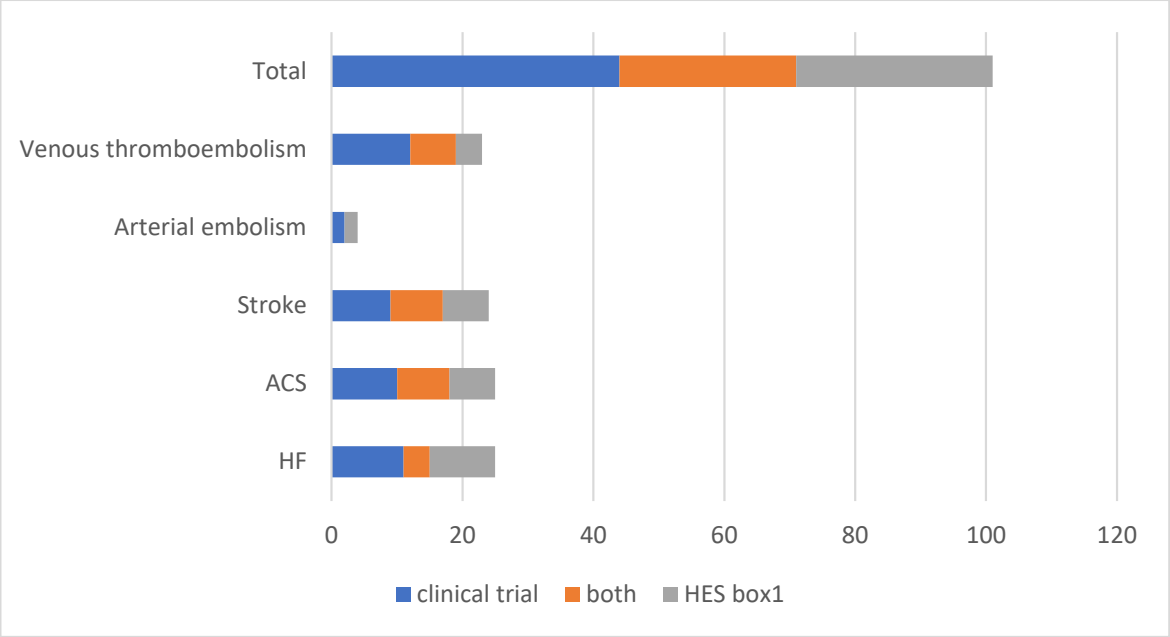
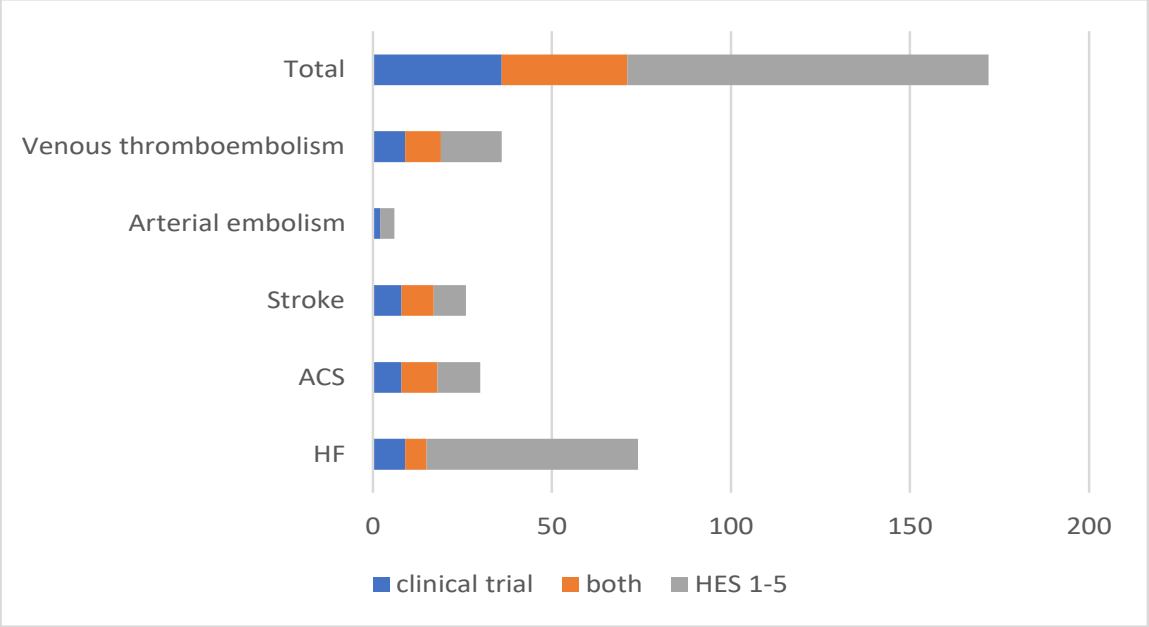


Figure 2-7– CTD data vs HES APC dataset (HES) diagnosis boxes 1-5



2.3.2 Triangulation comparison between Heart failure/ ACS clinical trial data, HES APC data and NICOR data

Comparison of the ACS and heart failure events was performed between the three datasets using a tolerance around dates of <= 2 weeks. Two triangulation comparisons were made between the three datasets; one using HES APC diagnosis box 1-5 data and then separately just using HES APC diagnosis box 1 data alone (**Table 2.5**).

Triangulation using HES APC diagnosis box 1 only, demonstrated an agreement between all three datasets of 11 (20.7%) of the cases. Agreement between PATCH trial data/ HES APC diagnosis box 1 only was 1 (1.9%) case; 1 (1.9%) PATCH trial data/NICOR and 9 (17.0%) NICOR/ HES APC diagnosis box 1 only. The following differences were seen with PATCH trial data alone 20 (37.7%), HES APC data diagnosis box 1 alone 8 (15.1%) and NICOR data alone 3 (5.7%). This is visualised in a Venn diagram of the data in **Figure 2.8**.

Triangulation using HES APC diagnosis box 1-5 shows a similar agreement between all three of the datasets with just 12 (11.4%) of the cases. Agreement between PATCH Trial data/ HES APC diagnosis box 1-5 data in 4 (3.8%) cases; 0 (0%) Trial data/NICOR and 11 (10.5%) NICOR/ HES APC diagnosis box 1-5 data is also similar. Events that occurred in a single data base comprised; PATCH trial data alone 17 (16.2%), HES APC diagnosis box data 1-5 alone 60 (57.1%) and NICOR data alone 1 (1.0%). This is visualised in a Venn diagram of the data in **Figure 2.9**.

That resulted in a PPV of 11/31 = 0.35 for events that appeared in both HES box 1 and NICOR, and 13/33 = 0.39 for events that appeared in either dataset. This equated to a sensitivity of 11/31 = 0.35 that appeared in both HES box 1 and NICOR and 13/33 = 0.39 for events that appeared in either dataset.

That resulted in a PPV of 12/74 = 0.16 for events that appeared in both HES boxes 1-5 and NICOR but only 16/88 = 0.18 for events that appeared in either dataset. This equated to a sensitivity of 12/29 = 0.41 that appear in both HES box 1-5 and NICOR and 16/32 = 0.50 for events that appear in either dataset.

Table 2-5: Three way comparison between CTD, HES APC dataset (HES) and NICOR

	Single source only			Two sources			All 3 sources
HF/ACS	CTD	HES	NICOR	CTD& HES	CTD& NICOR	HES& NICOR	CTD& HES& NICOR
HES box 1 alone	20	8	3	1	1	9	11
HES box 1-5	17	60	1	4	0	11	12

Figure 2-8: Venn Diagram of ACS/ Heart failure CTD, NICOR data and HES APC dataset Diagnosis box 1 data alone

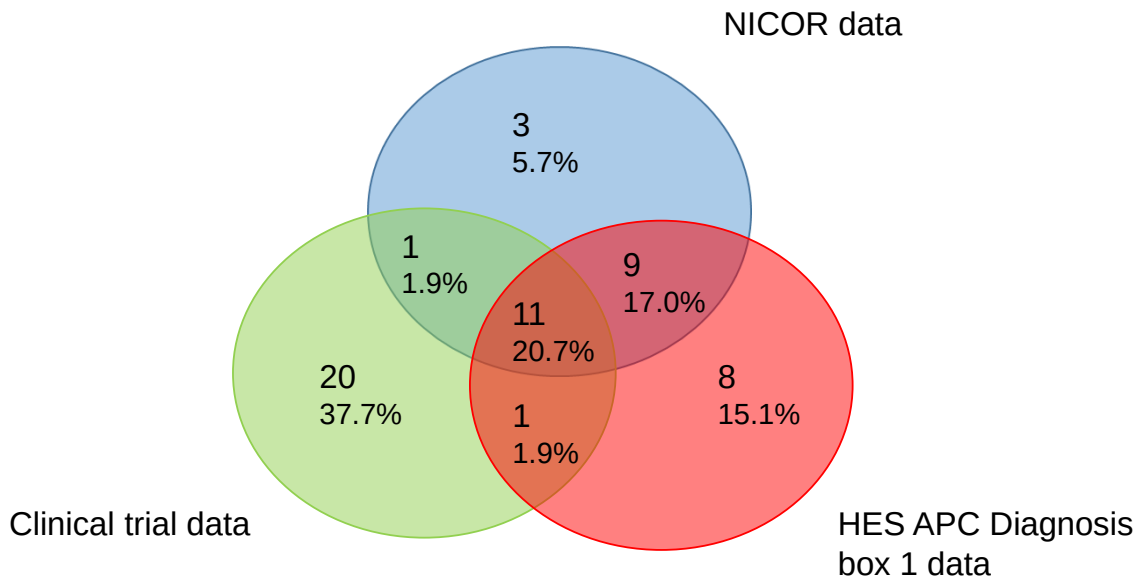
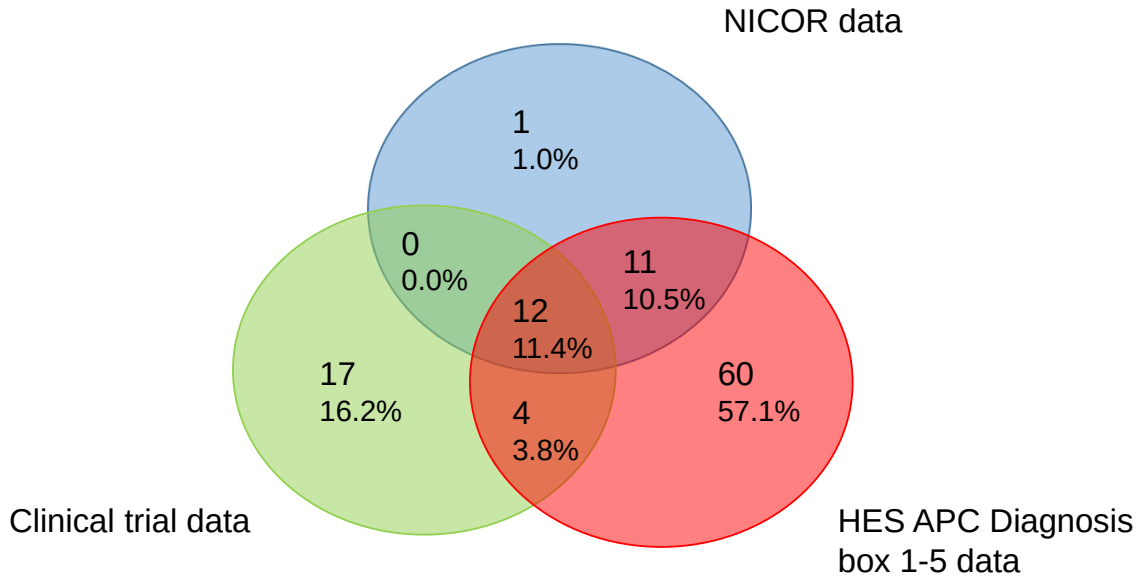


Figure 2-9: Venn Diagram of ACS/ Heart failure CTD, NICOR data and HES APC dataset Diagnosis box 1-5 data



2.3.3 Qualitative review of event descriptions

2.3.3.1 ACS/Heart failure

All events were then reviewed clinically to ascertain differences between the datasets. There was a marked difference in overlap and number of events within the Venn diagram if HES APC diagnosis box 1-5 or diagnosis box 1 only were used. HES APC only events data are the most common when diagnosis box 1-5 is used whereas trial events become most numerous when diagnosis box 1 only is used. This was primarily due to heart failure codes being often repeated in future events following diagnosis. When clinically reviewed most of these admission events were due to other health conditions but heart failure is still included in diagnosis box 1-5.

Clinical review of all the HES APC only events demonstrated that of the HES diagnostic box 1 events, 3 (2 heart failure/ 1 ACS) matched events recorded in the trial database that had not reached the criterion for trial CVS endpoint. Using the broader HES diagnoses 1-5 data, there were 6 events (4 heart failure/ 2 ACS) that appeared within the clinical trial data but had not reached the criterion for trial CVS endpoint.

On review of all the HES events alone demonstrated likely high false positive rate when all five diagnosis boxes are used with only 3/60 considered possible events. A possible event was defined as appropriate code, no repetition from previous events and no trial case report form to demonstrate an alternative diagnosis.

Clinical review of ACS/ Heart failure trial data alone showed that 15/17 events occurred solely outside of hospital so could never have been picked up by either HES APC or NICOR data. These were either managed in the community or in outpatient setting or out of hospital sudden deaths coded within the trial as ACS/ Heart failure. The other 2 trial events, not found in the HES APC, which were admitted to hospital either demonstrated perhaps poor coding within the HES APC data or were not found within the HES dataset. Where there was no record within HES APC data this may imply either linkage problems between the trial and HES APC dataset or the event did not get uploaded to HES APC.

There was 1 NICOR only event when diagnosis box 1-5 is used in HES APC. This changes to 3 events when diagnosis box 1 is only used within HES APC data. Of the

three events 1 event when reviewed had a CRF within the trial and the event did not meet trial criteria. The other two events cannot be confirmed as no CRF is present to compare to. When NICOR/HES APC diagnosis box 1 only events were reviewed 8/9 events have no CRF to suggest alternative diagnosis. These 8 could be possible events within the trial as there is no evidence for alternative diagnosis and NICOR clinical detail suggest either an ACS or heart failure event. The trial treatment of the participants was reviewed for all these 8 events. All of the 8 events were found to be within the control arm of the trial. It was also noted that this would not have changed the outcome of cardiovascular event analysis of the trial if they were included as in fact they would have strengthened the case for no difference between LHRHa and oestradiol patches.

2.3.3.2 Stroke/Venous Thromboembolism/Arterial embolism

Clinical review of trial data alone showed that 11/23 stroke/venous thromboembolism and arterial embolism events were outside of hospital so would never have been picked up by the HES APC. These were either managed in the community or in outpatient setting or out of hospital sudden deaths coded as one these cardiovascular events within the trial. There were 8 events that either did not have a HES APC event or did not have the appropriate IDC-10 code within diagnosis box 1-5.

As previously there was a marked difference when diagnosis box 1-5 or diagnosis box 1 only was used within the HES APC data. Two of the HES APC only events matched clinical trial CRFs but did not meet the defined clinical trial criteria for an event. When the thromboembolic stroke/ venous thromboembolism and arterial embolism HES APC diagnosis box 1-5 only events were reviewed 12/30 events were considered possible events. A possible event was defined as appropriate code, no repetition from previous events and no trial case report form to demonstrate an alternative diagnosis.

2.3.3.3 Review of follow-up in HES/ NICOR only events

Another consideration for those HES/ NICOR events that do not appear within the trial data is that participants in this defined cohort may be lost to follow-up or we are awaiting further detail as the event occurred following the last follow-up. When this is reviewed within this cohort approximately a third of events that possibly could have occurred in the EHR data were after the last trial follow-up which would give an explanation why they do not have a trial CRF for that event yet. In the events which were considered possible true events on the EHR datasets median follow-up time was over 3 years and for those events that were within NICOR and HES APC data had a median follow-up of over 3.6 years. This shows within this cohort that over half of the events were quite a few years following randomization which may possibly affect the degree of reporting at the site.

2.3.4 Review of inpatient hospital events only

On clinical review of all trial cardiovascular events, 26 out of 71 events occurred outside hospital so may not reasonably be expected to have been picked up by the either HES APC or NICOR data. These were either managed in the community or in an outpatient setting (n=21) or entailed out of hospital (sudden) deaths (n= 5). Therefore, these were removed from the trial events for a further comparison leaving a total of 45 events; 7 of episodes of heart failure, 11 ACS, 11 thromboembolic strokes, 14 venous thromboembolism and 2 arterial emboli. Summaries figures are described below to show the difference in comparison in **Figures 2.10 - 2.13**.

Therefore the PPV and sensitivities were repeated to confirm the difference if these events were removed from the analysis.

Considering the clinical trial data as the gold standard for all cardiovascular events, for events detected in HES diagnosis box 1, this translated to a PPV of 0.47 and a sensitivity of 0.60. Using HES diagnostic boxes 1-5 however this dropped to a PPV of 0.26 although with improved sensitivity at 0.78.

The same analysis was also carried out for ACS/HF only with HES and NICOR as previously. This resulted in a PPV of 12/74 = 0.16 for events that appeared in both HES boxes 1-5 and NICOR but only 16/88 = 0.18 for events that appeared in either

dataset. This equated to a sensitivity of 12/14 = 0.86 that appear in both HES box 1-5 and NICOR and 16/18 = 0.89 for events that appear in either dataset.

Figure 2-10: CTD data vs HES APC dataset (HES) diagnosis Box 1 for inpatient events only

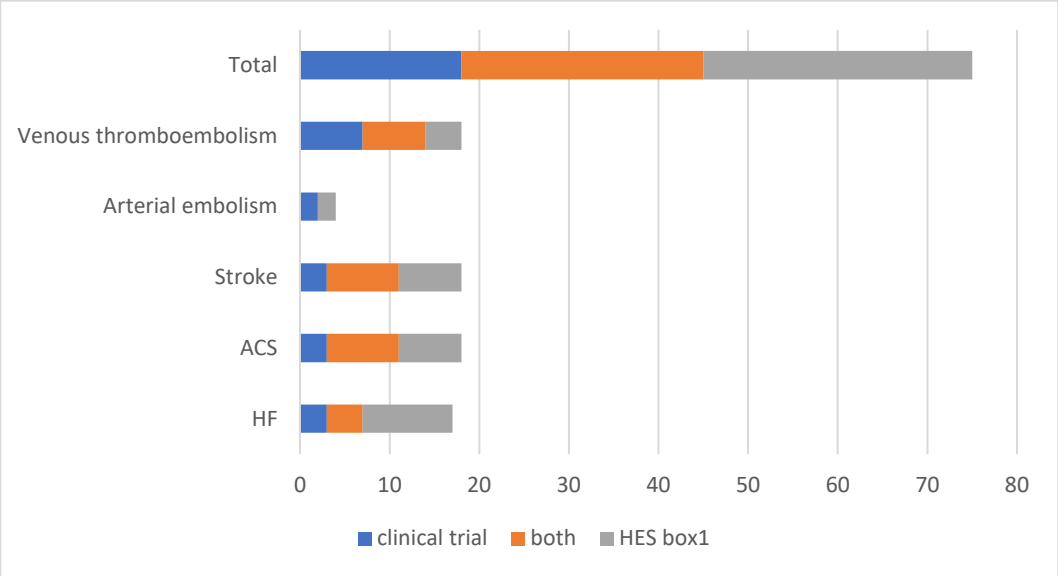


Figure 2-11: CTD data vs HES APC dataset (HES) diagnosis Box 1-5 for inpatient events only

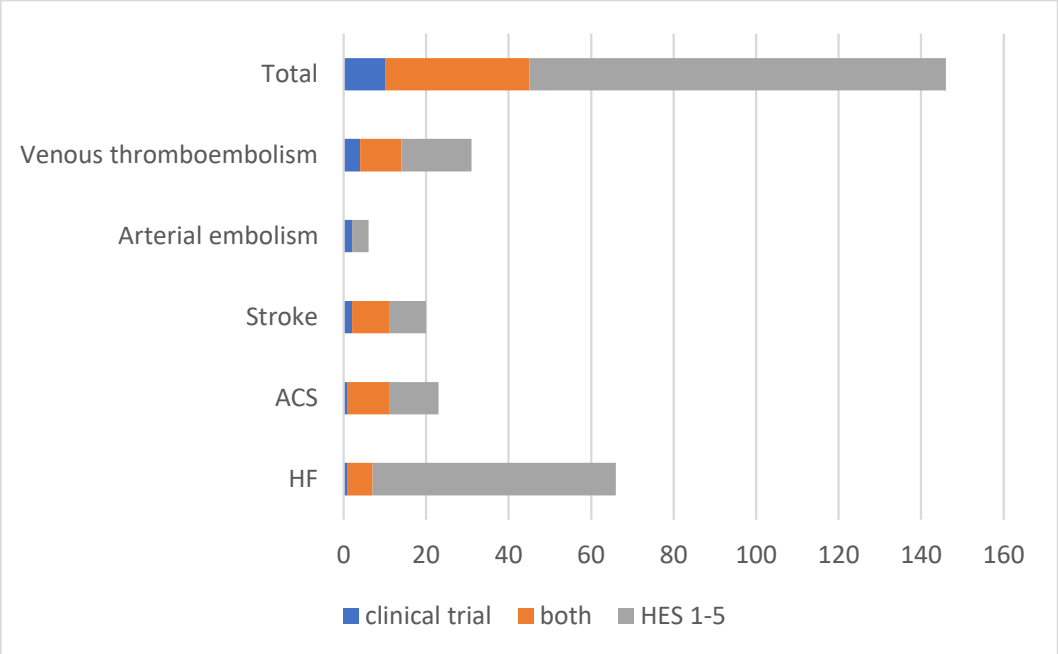


Figure 2-12: Venn Diagram of ACS/ Heart failure CTD, NICOR data and HES APC Diagnosis box 1 data alone for inpatient events only

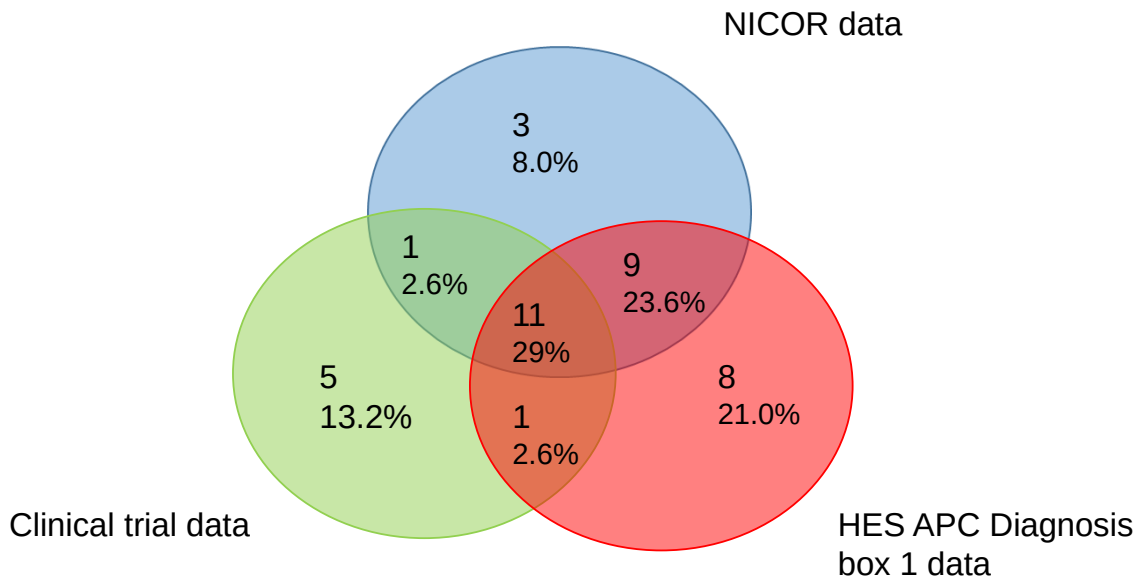
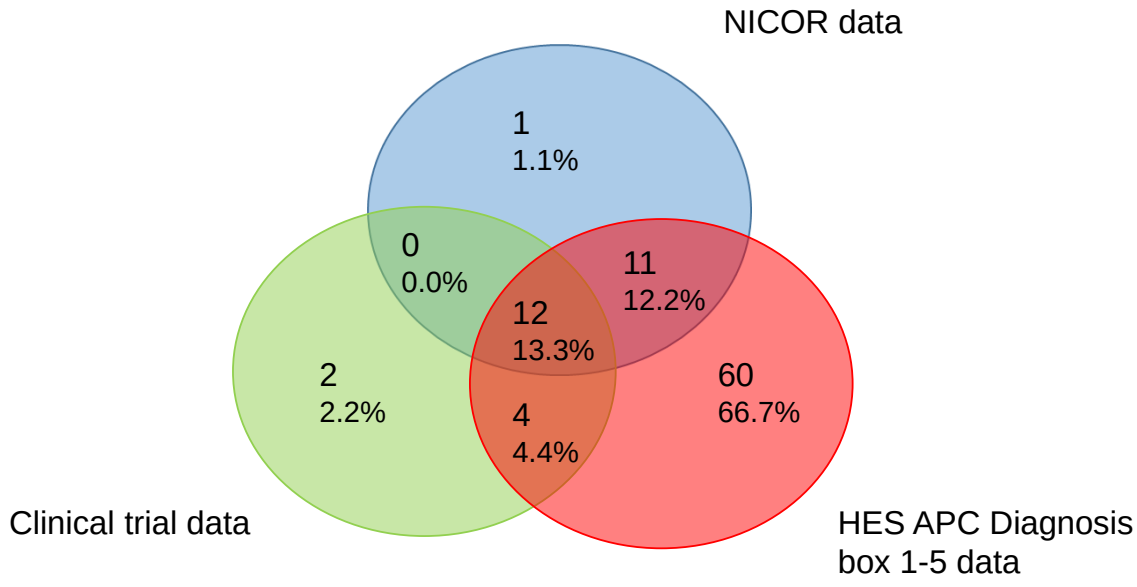


Figure 2-13: Venn Diagram of ACS/ Heart failure CTD, NICOR data and HES APC Diagnosis box 1-5 data for inpatient events only



2.3.5 HES A&E analysis

In total when HES A&E data was reviewed there were 4,362 separate hospital attendances. When the diagnosis box coding was reviewed there were 1,104 (25%) events with no coding (blank cell). There were 548 (13%) events where the code was diagnosis “38” – Diagnosis non classifiable. These could conceivably be events that do not fit into any of the diagnosis boxes but this seems unlikely as the data dictionary is comprehensive covering all body systems. Therefore, it is more likely that this has been poorly filled in or A&E data cannot give a clear diagnosis. Both of which mean that that event data is unusable (similar to blank) as we do not have any information about why the patient came in. This means that the rest of the analysis is difficult to make firm conclusions from as 38% of the data has no diagnosis.

From the data that we have with diagnosis codes there is a similar pattern with only 561 (13%) of codes following the A&E data dictionary 6 digit coding. Also only 60 (1.4%) events used ICD-10 coding similar to HES APC data. When all potential codes are reviewed clinically to see if they could be used for a diagnosis then there were 100 potential events within the time period. 11 match CRF Cardiovascular endpoints and 25 any CRFs in the trial. Of the events with matching trial CVS endpoint events -3/15 heart failure, 4/18 ACS, 3/17 cerebrovascular event, 1/19 venous thromboembolism. 77 events were just present in the A&E dataset and had no corresponding trial CRF. These events cannot be confirmed as true events due to poor nature of coding within this dataset. One event which would not have been picked up by HES IP was picked up by A&E data for right leg DVT- coded appropriately.

2.4 Discussion

This study analysed data including 8 years of cardiovascular events that occurred within a cohort of patients enrolled in the PATCH trial, incorporating trial data from PATCH and corresponding EHR from HES and 2 NICOR audits (MINAP and Heart failure). This statistical method of concordance was a PPV and sensitivity analysis. The disparity between trial only event and no registry recorded event (sensitivity) and

registry recorded event but no matching trial event (PPV) with trial data considered as the ‘gold standard’.

Overall there was relatively poor concordance between the three datasets. This seems to be particularly when NICOR and HES APC Diagnosis box 1 for ACS/ Heart failure events were analysed at best a PPV value of 0.39 and sensitivity 0.39 when all possible ACS/ Heart failure events are considered. The sensitivity only slightly improved when diagnosis boxes 1-5 (i.e. a broader more permissive approach) are used in HES APC with a sensitivity of 0.50 to the detriment of the PPV 0.18. In the analysis of all cardiovascular events there was also a similar picture when diagnosis box 1 was used (PPV 0.47 and Sensitivity 0.38) and diagnosis box 1-5 (PPV 0.26 and sensitivity 0.49).

However, the Venn diagram and statistical analyses under appreciate the correlation when events are interrogated in more detail. In this analysis it is clear that if trial events that did not happen within an inpatient setting were removed then the sensitivity of EHR records increase dramatically (0.60 for all cardiovascular events for diagnosis box 1/ 0.79 diagnosis box 1-5). This can be seen most clearly within ACS/HF when HES diagnosis box 1-5 is used the sensitivity rises to 0.89, however PPV does not change. This would be more in keeping with cardiovascular trials that have compared trial data with cardiovascular events previously with WOSCOP analysis (98) and more recently with the ASCEND (A Study of Cardiovascular Events in Diabetes) trial (NCT00135226). When the trial data was compared with HES APC and death data for nonfatal myocardial infarction, ischemic stroke, TIA, or vascular death, excluding haemorrhagic stroke where there was a sensitivity 0.72. (122)

It is clear that there is a trade-off between sensitivity and PPV in looking at events in greater detail; even though you might get a few more events from using 1-5 diagnosis boxes in HES APC data this means that the false positive rate particularly of a diagnosis of heart failure goes up markedly as this is repeated in multiple admissions where heart failure may have not been the cause of the admission. This is because important historical or ongoing medical conditions often are included in diagnosis boxes for admissions as they contribute to the care the patient receive and therefore the hospital should receive financial compensation for the management of the more complex patient (37). As HES APC data has been designed for financial compensation

this demonstrates appropriate coding for its purpose. However, for the purpose of trials it may mean that you can only use diagnosis code 1 if you want to know why a patient was admitted to hospital at that event, without collating too many false positives. This is in line with previous trial use of HES data (79).

It is interesting that for serious acute events like ACS and stroke this was not such a factor and there may be an argument for inclusion of 5 diagnosis boxes or more in certain circumstances within trials using EHR data. The cut-off point of diagnosis box 5 was determined upon initial preliminary analysis and as the median total of diagnoses boxes completed was approximately 5. This could infer a limitation to this study and more diagnosis boxes could have been included to give more information on an appropriate cut-off and try and include all events. This formal assessment was not within the scope of this study but could be part of future work.

The definition of trial data and NICOR data are very important in considering lack of concordance. NICOR, especially more recently, have only used definitions based on HES for certain codes in diagnosis box 1. This reflects the very smaller number of NICOR alone events and good concordance with HES. However, unless trial outcomes match NICOR definitions exactly this may mean missing events if the trial used only NICOR to collect long-term data on trial participants.

Within the data collected through the trial procedures, the event can happen anywhere i.e. not just in hospital. If this event results in a death then electronic health records could collect this data as part of ONS death certification and this would be an important additional of EHR out with this current study and future studies. If it is an ACS event then these data might potentially be available via other NICOR audits not used in this study, such as the NICOR Adult cardiac surgery or Adult coronary percutaneous intervention audits. However, if these events are treated all in the outpatient setting then then NICOR and HES APC data would not pick this up.

The inclusion of a fourth data set, namely GP data, may solve this problem; however at time of investigation there was no single equivalent GP dataset set that covers the whole of England. This is an important distinction for the case of RCTs as opposed to cohort studies that by design will include only patients where data is available via the GP CPRD (that currently incorporates just s subset of UK population). NHS Digital have recently started to make available a pan England GP dataset and this could

definitely change the concordance of these datasets markedly as previously shown by the Herrett *et al.* in their triangulation paper (79). The importance of the GP dataset has recently been supported by the 'CVD-COVID-UK' Consortium in their publication. They reviewed cerebrovascular events and acute coronary events with COVID datasets. In their publication they compared data from HES and national GP data via NHS Digital. The GP data encompassed 96% of the English population which was far superior to previous GP dataset providers. This demonstrated that 30% of TIAs/Strokes are only coded in primary care alone and 12% in the death registry. For MI 8% were also only seen within primary care records and 12% in the death register (123).

There were a number of events that were seen within both NICOR and HES but had no corresponding case report form within the trial. Due to the nature of NICOR requiring extensive information about the event, it is likely these are true events but have not been submitted to the trial by the site. We were unable to confirm these events with the site as our data sharing agreement did not allow us to re-identify patients or use this data for study outcome. When the pseudo-anonymised trial data was analysed however it was notable that all these events were for participants on the control arm, and a third were after the last follow-up on the trial, often many years after randomization. This may reflect what has been reported in other trials in that there is an attrition of data over time which can be dependent on the allocated arm, demographic factors and disease state. (22, 23, 124) This strongly supports the use of EHR data in the long-term to truly reflect the events in both the control and research arm. As an important aside it is important to note that if anything this reinforces the cardiovascular safety of the experimental approach (tE2) within the PATCH trial.

In the primary analysis, where trial data is compared with HES data for other and all events, convey a similar message showing that HES APC is less specific but may be more sensitive for hospital events then trial data. This is demonstrated in that there are more possible events with the HES APC-only data then in the ACS/Heart failure comparison. However due to the nature of these conditions, venous thromboembolism, thromboembolic stroke and arterial embolism, many of these events are managed outside of inpatient hospital episode. This would mean possible events would not be picked up compared to ACS or acute heart failure where most

are seen as an inpatient. For these events there would need to be another data source if used for long-term follow-up which again could be achieved by a national GP dataset.

To solve this problem A&E HES data was analysed to see if this could be used as a secondary source to collect events of this nature like TIA, DVT, and Pulmonary embolism (PE). This is because many of these conditions may be treated in this setting (A&E attendance but never admitted to hospital). However, on review of the A&E HES data the diagnosis box was very poorly filled either blank, code for diagnosis unknown or not in a format that would be appropriately for statistical interrogation. When A&E HES events with diagnosis codes were compared with trial data there was very little similarity and numerous trial events completely missed with A&E data. There was one event for deep vein thrombosis which was picked up by A&E HES and not HES APC data. However, this is not sufficient to consider this an appropriate dataset to be used in this situation. This reflects the purpose of the A&E HES data previously where diagnosis was not subject to different funding so incentive for this to be filled in was not present especially with the increased volume of attendances compared to HES APC data. A&E data has been successfully used in other studies like 'routes to diagnosis' work by NCRAS (125) which did not rely on fields having a high fill rate or concordance to their data dictionary. The HES A&E Dataset is now being superseded by the Emergency Care Dataset. This has been gradually replacing the A&E dataset over the last couple of years and hopes with improved SNOMED coding that they will get better data capture for diagnosis in the future (126). However, as many trials would be using historical data the recommendation is not to use A&E HES data for our proposed purpose.

There are limitations to this study. The actual event rate within PATCH for cardiovascular events is relatively small. This limits extensive statistical analysis of these events; only descriptive analysis is appropriate. It may also mean statements about true concordance between the datasets are not generalizable. However, the final stage of the PATCH trial has its primary outcome as overall survival and the hypothesis at the outset was for there to be limited number of cardiovascular events i.e. confirming the safety of transdermal oestradiol in this setting. This would be similar with many other cancer trials where there would be a cardiovascular risk but in the population studied the event rate might be relatively low. Therefore, this study reflects how many other oncology trials would use EHR data in this setting. There is also a

strength in low numbers in that each event can be reviewed and scrutinized to see if they are events from the narrative or coding. This is especially true of review of the 1-5 diagnosis box versus 1 only where you can demonstrate that there are many different reasons for admission and explain the false positive rate.

We have also studied only a subset of the PATCH population including patients recruited from 2010 to 2018 and then only from England. Findings are then only strictly relevant to the English participants as event though the registries are likely to be at a similar standard the other nations registries have not been analysed in this project. The linkage between NICOR and NHS Digital would also have been more appropriate for this study to make sure we got all appropriate NICOR events. However, this is also a strength of the study as it shows what data is retrieved if you applied to only one of these registries alone rather than both together. This is important to know as each access incurs significant costs; if you could do without one of them then that would be to the benefit of the trial community in the future.

Lastly some may argue that not re-identifying the patients to check if EHR events were true is also a limitation. However, the data sharing agreement meant that this was not possible due to consent concerns. This was also a conscious decision by the trial team as this was a subset of patients who were never supposed to inform trial outcome and the protocol does not state that outcomes could be determined by registry data. If some patients' events were gathered from EHR data it was thought that this could possibly bias the final outcome. It also meant that data collected by the registries did not need to be held by the trial for the regulatory specified time which could put the trial at high financial risk as data agreements would need to be extended at a high cost to the trial.

This study demonstrates that if you are to use EHR data alone or for long-term follow-up then the protocol trial outcome must reflect that and meet the definition of the EHR records that you are using. If the trial event definition is slightly different then EHR data could be used more as a trigger to go back to sites to see if a true event had occurred and could allow for events not to be missed especially in long-term follow-up. This study reflects that no record is perfect and to get a 'true' reflection of the events then all three may be needed in the case of ACS and Heart failure and HES/ trial data with the other events. This study raises the question of 'how true events are defined?' Trial

data has the benefit of information being collect from many clinical settings and also being scrutinised by site and trial physicians. Trial data however may miss events if not reported to the trial team. HES APC should collect nearly all hospital events if the right codes are used. It may need a clinical review at site to determine if the event matches the protocol definition as hospital coding may not be as rigorous as trial collected data. NICOR may miss certain events due to its reliance on data imputation at site and also ongoing strict coding criteria in inclusion of the event. However, NICOR gives vast array of data that can be used to confirm diagnosis, when compared to HES, that may be essential if you were just to use EHR data for long-term follow-up. In essence the truth in trial data is how you define your endpoint at the start within your protocol. It is essential when you are presenting your work that the outcome definition is clear and how you collected it especially which code and database if using EHR data. For the future if EHR will be used for cardiovascular long-term outcomes then the use of the consort extension for clinical trials is essential (127) and also a clear definition of what is defined as a ACS event within the EHR knowing that this definition has changed with time. For ongoing trials this should be based on NICOR definitions or another certain standard to make sure that trials are comparable.

This study demonstrates certain challenges of using electronic data that make using these resources more difficult for trialists. Application for data is at present a very complex process and if you have not set up the trial with appropriate consent and protocol wording this makes it even more difficult and sometimes impossible to get long-term data. From the study we can see that the more data sources you have the more reliable you feel the data is to give a true reflection on event rate. However, the cost increases dramatically as you add in different datasets particularly with NICOR. The cost for getting repeat extractions or flagging patients also mounts up so that if you follow participants up over time with multiple extractions the amount of money can dramatically increase. Lastly the data contracts are also very precise on how long you can have the data and often the longest you can hold a contract is a year or up to 3 years with NHS Digital. To renew the contract or keep hold of the data also comes at a cost and also resubmission of paperwork which may also be an unforeseen cost in the time it takes to fill in the application. At present many of these problems put trials off using this data. However a lot of these can be managed by making sure that there is a good relationship with the data source prior to starting the trial to make sure that

appropriate paper work and project design is compatible with what they can provide. This can also give an idea of the cost so that you can integrate this into grants or see if data collection through these databases is reasonable.

2.5 Conclusion

Comparing cardiovascular events as captured by the PATCH CRFs, HES data and NICOR audits show a surprising variability and relatively poor concordance between datasets when all events are assessed. However the sensitivity of inpatient cardiovascular events is high, especially ACS/Heart failure events, when compared to trial data. Clearly at this time, no single dataset can be recommended as an alternative to trial CRFs, and yet these data also raise questions about what we consider the gold standard when reporting clinical trial outcomes. In planning to use EHR in this way, significant efforts must be made in defining specific outcomes such that the intended data might be collected but it is hard not to conclude that numerous data sources will be required to get as close as possible to the ‘true’ data. The human and financial costs of such EHR data access and retention are considerable; an important consideration and certainly not representing a more efficient approach.

3 Feasibility of long-term follow-up of outcome data within the Add-Aspirin trial using EHR data

3.1 Introduction

3.1.1 Long-term follow-up

The importance of long-term follow-up of individuals participating in oncology trials is increasingly recognised, though there are both financial and resource implications. Long-term follow-up is required as treatments become more successful and participants are surviving longer, and also because some side effects occur many years after treatment, for example following radiotherapy (128). Additionally, certain cancers recur after many years of dormancy. Metastatic bone recurrence in breast cancer can occur 10-20 years after initial radical treatment for hormone sensitive disease in the post-menopausal setting (59, 129). Prostate cancer can also recur many years after initial treatment with biochemical progression and then definitive disease progression on radiological scans (130, 131). These are not the only cancer types where this occurs but they represent two of the most common cancers, and approximately a third of all new cancers in the UK (6). This is reflected in clinical trial activity where breast and prostate cancer trials were the 2nd and 5th most common trials registered by cancer type between 2000-2018 (132).

The number of oncology trials requiring long-term follow-up of participants has been increasing (133). This puts a heavy burden on the research community. This is either with the CTU or the NHS. Both need to continue to employ administrative staff and data managers to collect and process data over many years. Long-term follow-up has previously been shown to be one of the most costly aspects of any trial (134). In NHS clinics for common cancers, e.g. breast/ prostate/colorectal cancer, the face-to-face follow-up of patients has been reduced (135). This decreases the burden on a heavily oversubscribed oncology outpatient system to allow for new patients and patients on active treatment to have readily available appointments when needed. There has been a change to a more patient led reporting of symptoms or surveillance system based

on routine radiological tests and blood tests to pick up recurrence rather than clinic appointments (135). This approach has been shown to be just as effective as routine follow-up and an increasing number of NHS services are adopting this new way of working. However, some patients participating in clinical trials require extra follow-up appointments compared to non-trial patients to collect the relevant data which in turn incurs extra cost and human resource (136).

Kilburn *et al.* demonstrated the issue in early breast cancer trials. In their study they contacted CTUs to review case report forms and what information was collected during long-term follow-up. It demonstrated the burden on NHS sites with 76.5% of trials asking for the patient to be seen in person (133). In a questionnaire to NHS sites, that were collecting data for the trials, common themes were identified to provide evidence for developing a paired back long-term follow-up form which they present in the paper. Important common themes were a lack of personnel, resources and room for storage to cover long-term follow-up; CRFs were too lengthy and also the CTUs often asked for a large amount of data in a very short space of time. All of this is on the background of funding for trial work prioritising recruitment and not follow-up at NHS sites.

The traditional methods of follow-up in RCTs, as stated in the thesis introduction, risk the loss of participants due to drop out. This may be more apparent in particular patient groups which may lead to potential bias over long periods of time (22, 137, 138). Therefore, considering all these factors there needs to be new ways of collecting long-term data. One method to reduce burdens on sites is, as Kilburn describes, is to reduce number of CRFs; amount of information required on each CRF and the number and frequency of trial visits. However this does not solve the issue of participant consultations, even if by telephone, as these are still the most costly method of follow-up (136).

Other methods that have been considered are questionnaires which are sent out to patients directly and have been used in a number of trials. Examples include in the CA125 and Ultrasound in Detecting Ovarian Cancer in Postmenopausal Women trial (UKCTOCS; NCT00058032) where they were used to collect information about the side effects of treatment. It was argued that this provides a more accurate side effect profile then from the traditional clinician led follow-up (139). This type of data could also be obtained in the form of electronic captured PROMS which also has growing

evidence of validity (140). However, for outcomes such as cancer diagnosis the results from validation exercises of this method have been shown to be mixed and possibly not as effective as traditional or other methods of long-term follow-up (69, 141-143). Also, for mortality data it would not be appropriate to ask relatives for cause and time of death.

A more favoured approach to long-term follow-up is the use of EHR or routine data sources. There are many examples of this in trials either supporting original analyses or even demonstrating differing results or effects on interventions (128). The main outcomes trials use EHR for are death. New cancer diagnoses and cardiovascular outcome are also collected in this way but primarily in Scandinavian countries and the USA (128). The UK has more experience than other countries in the rest of Europe and influential trials have used EHR for mortality and cancer outcomes (60, 144-148). However, many trials have used this resource in the UK without evaluating if the data source is good enough compared to usual trial methods. Mortality data has been evaluated by a number of trials and these publications have been supportive of the use of EHR mortality data in the UK (142, 149-152). However even some validity studies have demonstrated that there is enough difference between mortality data to give different survival outcomes (66). This specific study was older than the more recent validations possibly demonstrating improvement in the accuracy and completeness of routine datasets over time. As with all validations of data sources it is important to continue to assess as they were not designed for trial purposes so may change over time.

3.1.2 Cancer registry validity in long-term follow-up of oncology trials

Using national registries for long-term trial follow-up has been successfully used in the past in large studies (60, 144-147) but there are still concerns over the completeness and accuracy of information of EHR versus traditional active follow-up. The level of accuracy of information will often vary depending on different outcomes/ database that are being interrogated. NCRAS data has been shown to have national coverage and high level of uptake, linkage to trial participants and accuracy in the primary diagnosis of cancer based on NCRAS analysis and external verification (67, 75, 142). There are

some acknowledged limitations on the accuracy of data relating to the staging of cancers which previously has differed between trial and registry data. In the past this has been attributed to high levels of missing staging data (22-44%) in the EHR but this is from data over 10 years old (151, 153). Merrel *et al.* in their more recent comparison of prostate cancer staging again demonstrated variable comparability with Gleason score having a good agreement (K= 0.90) but TNM score much less accurate (T 0.35 N 0.51 M 0.58). This is most likely accounted for by the fact that registry staging is based on many sources with interpretation needed to create a 'best' staging (67, 151). In a recent analysis of trial data versus cancer registration the UKCTOCS study has demonstrated when the cancer registration data was assessed against confirmed cases within the study between 2001 and 2014 that there was a sensitivity of 85.0% and a specificity of 94%. They concluded that cancer registration data alone would not have found all the relevant cancers in this screening trial (154).

For an adjuvant oncology trial, disease free survival is often the primary outcome measure or an important secondary outcome. However, historically there has been a lack of accurate recording of recurrence of cancers in registry data. NCRAS have recently published very low figures for recurrence data even with the development of new datasets namely 'Cancer outcomes and services dataset' (75, 155). This was evaluated in detail when the "trial of accelerated adjuvant chemotherapy with capecitabine in early breast cancer (TACT2; NCT00301925) trial" data was evaluated against NCRAS data (151, 156). The comparison demonstrated that concordance was generally good for demographics and death. Recurrence of cancer had a poor concordance with only 63% of distant recurrences and 70% of local recurrences being picked up in the NCRAS data. Again, this study was performed before many of the new datasets for NCRAS had been set up so this may have changed with time. The same group are aiming to do similar comparisons using a number of breast trials with more up to date data (153). For English cancer registry data to be more widely used in trials there needs to be on going assessment in multiple different tumour types especially reviewing the validity of staging and recurrence data in the datasets.

3.1.3 Add-Aspirin trial

Add-Aspirin (NCT02804815) is a phase III, multi-centre, double-blind, placebo-controlled randomised trial with four parallel cohorts. Each of the four cohorts are tumour site-specific. The primary aim is to assess whether regular aspirin use after standard therapy prevents recurrence and prolongs survival in patients who have undergone treatment for non-metastatic common solid tumours (breast, colorectal, gastro-oesophageal and prostate cancer). An overarching protocol facilitates a combined analysis of overall survival as a co-primary outcome measure as well as allowing individual site-specific analyses (157). All four tumour types have a separate defined primary outcome measure, evaluating disease recurrence and survival. It will recruit approximately 11,000 participants in the UK, Republic of Ireland and India. The design and outcomes of the study are summarised in **Figure 3.1**.

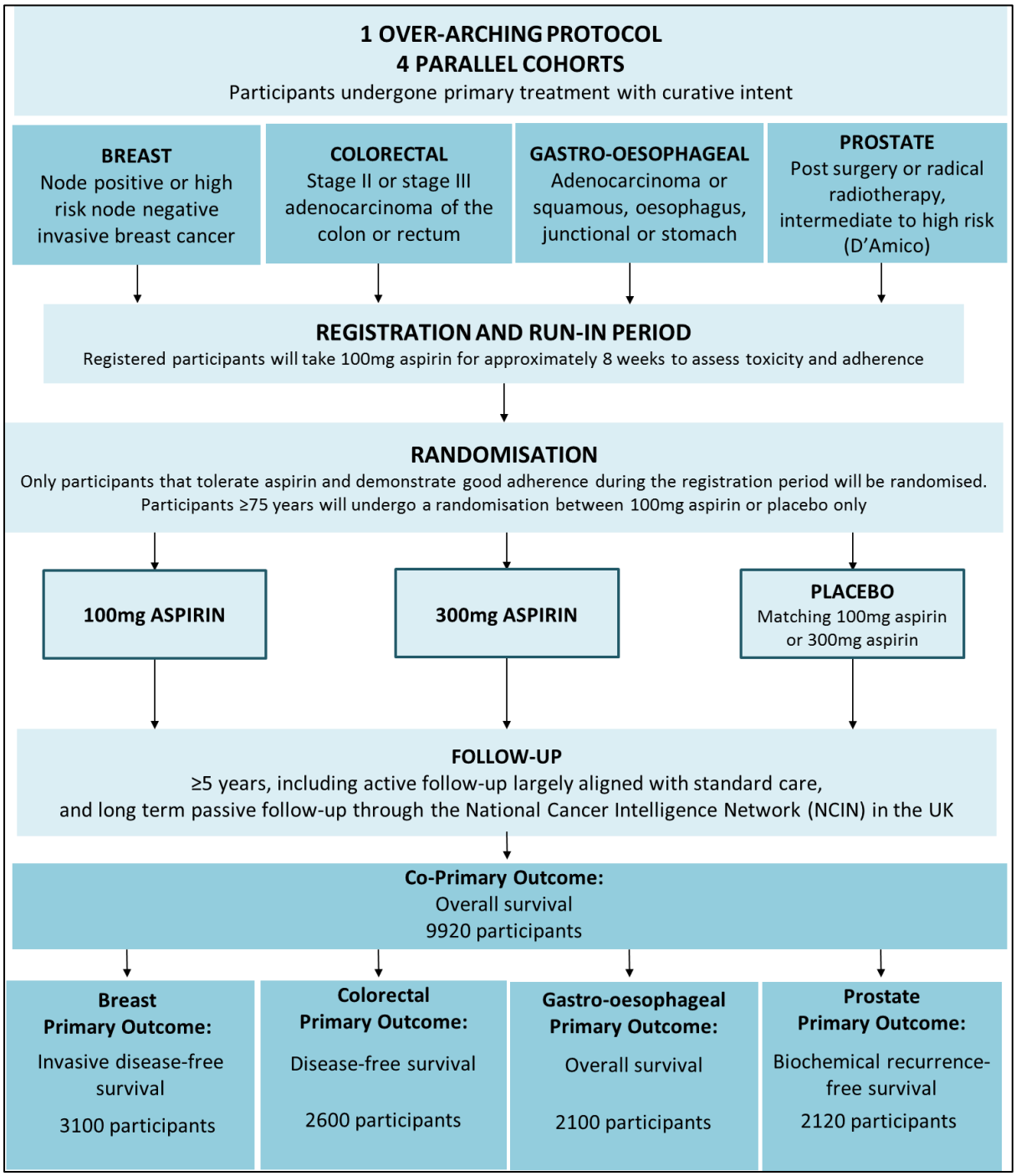
The data supporting the use of aspirin in secondary cancer prevention comes from pre-clinical, epidemiological and RCT data (157). In an analysis of 7 RCTs aspirin use was associated with a marked reduction in cancer death over 5 years of follow-up (HR=0.66, 95% CI 0.5-0.87, p=0.003) and an absolute reduction in 20-year risk of cancer death of 7% for those over 65 years (158). The effect was largest for adenocarcinomas (HR=0.53, 95% CI 0.35-0.81) and for gastro-intestinal cancers (HR=0.46, 95% CI 0.27-0.77). A subsequent publication involving 17,286 trial participants with a median of 6.5 years follow-up demonstrated that aspirin use was also associated with a decrease in the risk of developing any type of cancer with distant metastasis (HR 0.64, CI 0.48-0.84, p=0.001) and again particularly in adenocarcinomas (HR 0.54, CI 0.39-0.77, p=0.0007) (159).

Considering the tumour types included in the Add-Aspirin trial, cancer recurrence (following surgery or other similarly curative treatment) may not occur until many years after initial treatment; and the current evidence-base suggests that it will take a number of years of regular aspirin use before benefits may begin to emerge (158). Similarly, long-term follow-up is important in relation to other outcomes (such as adverse effects particularly gastrointestinal and other major haemorrhage), to ensure the risk/benefit profile can be holistically assessed. UK participants in the Add-Aspirin trial would normally only be followed up in the hospital setting for no more than 5 years but they

have provided consent for the linkage of routinely collected data to be used to augment trial data and provide long-term data after active follow-up has ended.

Add-Aspirin opened in October 2015 and recruitment continues. Some participants will no longer have regular hospital follow-up after 5 years and therefore the best and most efficient way of following these patients over time needs to be agreed. This methodology study will look at the most significant outcomes needed for the trial and if data from NCRAS could support or even replace long-term follow-up in participants from England (who represent approximately 85% of the UK participants in the Add-Aspirin trial). The feasibility study will concentrate on the trial data available up to April 2020 with registry data up to December 2019. The annual cost of data extraction will also be reviewed depending on the national registry data used. It will also describe gaps within the data where further work is needed or novel algorithms could be tested in the future to capture those events.

Figure 3-1: Add-Aspirin trial schema directly copied from Add-Aspirin protocol Version 5 (12 December 2016) (157)



3.2 Methodology

3.2.1 Objectives

The aim of this project was to assess the suitability and accessibility of routinely collected healthcare data in England for assessing outcomes in a large, multicentre, cancer clinical trial (the Add-Aspirin trial), and the potential for this data to ultimately replace long-term follow-up in the trial.

This was achieved through the following specific objectives:

i. To assess the feasibility and cost of obtaining timely clinical trial outcome data from NCRAS during early follow-up of Add-Aspirin trial participants.

ii. To assess the quality and completeness of this data (by comparison to data collected within the trial) for a range of different measures, including: cancer registration and trial outcomes such as cancer recurrence, death, second primary cancers, and SAEs as highlighted in the protocol listed as the primary and secondary outcome measures.

3.2.2 Ethics and data regulation law

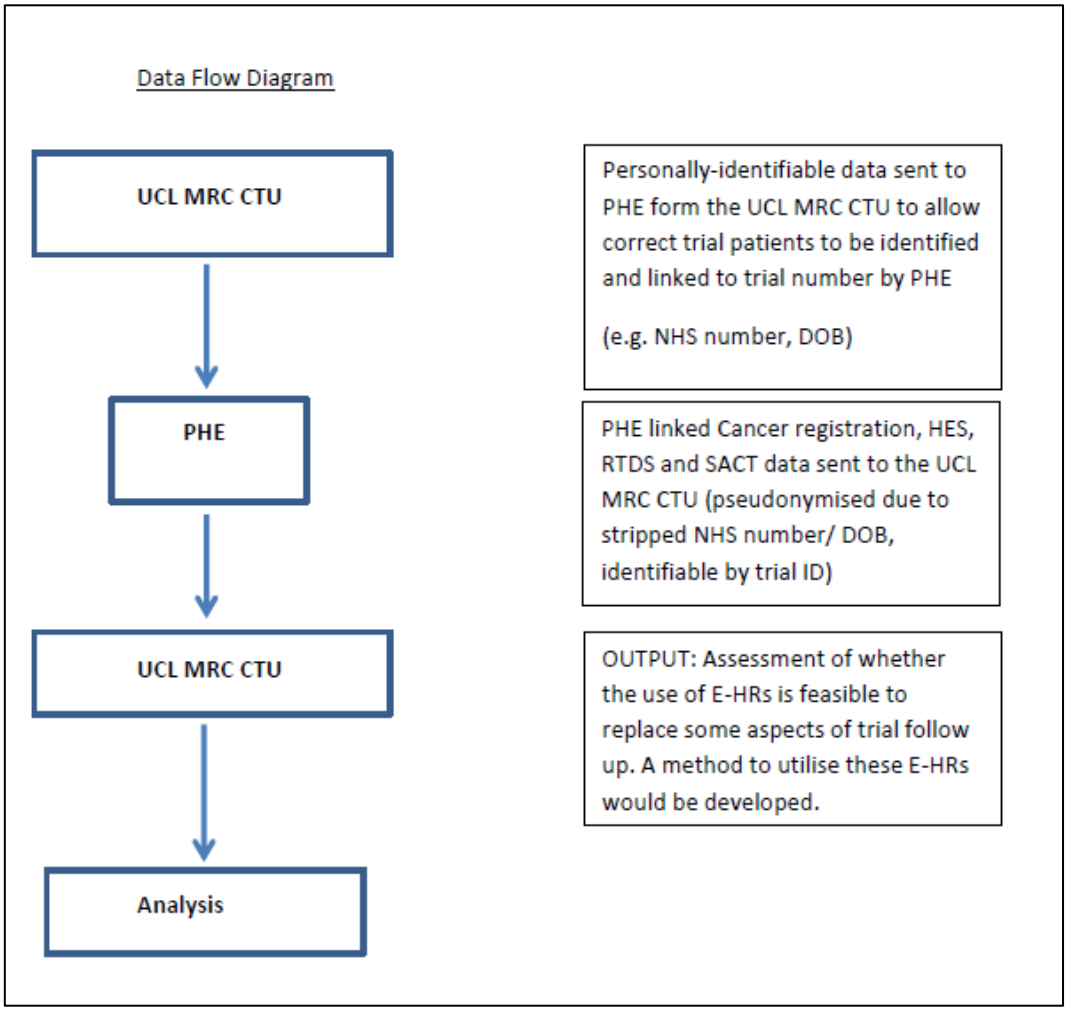
The Add-Aspirin trial was approved by the South Central – Oxford C research ethics committee (REC) and is part of the UK National Cancer Research Network (NCRN) portfolio. Patient consent forms, patient information sheets and trial protocol (which included this sub-study from the outset) have all been reviewed by a REC committee and have had MHRA approval. Details around results dissemination were further agreed with the IDMC, and supported by the trials steering committee (TSC), since it concerns release of data from the ongoing trial.

Trial participants had the option to provide their consent to allow their identifiable data to be used to obtain information about their health status from central registries (NHS Digital and NCRAS) at the time of providing consent to enter the trial. Data was only used from participants who had specifically provided their consent to this process. A transparency/ privacy notice was also written for participants as per data protection

law to describe exactly how their routine data would be used within the trial. This was made publicly available on the Add-Aspirin trial website and is attached in the thesis **Appendix A**.

The flow of data is described in **Figure 3.2**. All information transferred to NCRAS and then held by the MRC CTU was achieved within data governance laws and a risk assessment was performed via the data security and protection toolkit. At MRC CTU, the data was stored and analysed within the UCL data safe haven. The data safe haven has been certified to the ISO27001 information security standard and conforms to NHS Digital's Data Security and Protection Toolkit. It was built using a walled garden approach, where the data is stored, processed and managed within the security of the system, avoiding the complexity of assured end point encryption. Only authorized personnel were able to review and analyse the data. Patient identifiable data collected within the trial was stored within a separate database which is only accessible to senior members of the data management systems team and not the trial team. Identifiable data was only transferred between organisations in the first step of the process to allow identification of relevant individuals within the NCRAS datasets which was carried out by Mary Rauchenberger, head of data and management systems at the MRC CTU. Otherwise, data was pseudo-anonymized using unique trial ID numbers. No identifiable data was held within the project specific section of the data safe haven. No NCRAS data was removed from the data safe haven. Patients were allowed to be re-identified within the trial database by using their unique trial number to ascertain or confirm information from their trial site which had been discovered through the NCRAS routine data source and had not previously been reported on trial CRFs. Data was not shared with any third party. It was agreed that data would be published as per NCRAS guidelines. The process and timelines of the application to receipt of data from NCRAS has been summarised in the results section.

Figure 3-2: Data flow diagram from MRC CTU to Public Health England (PHE)



HES: Hospital episode statistics; RTDS: Radiotherapy Data Set; SACT: Systemic anti-cancer treatment data set; DOB: Date of birth, PHE: Public Health England

3.2.3 Datasets for comparison and censorship dates

3.2.3.1 Trial data

This sub-study included all Add-Aspirin participants randomised in England during the first 3 years of recruitment in the trial (Dec 2015 – Dec 2018) who had provided consent for this part of the study. Participants from India, Wales, Scotland and Northern Ireland make up a small proportion of the study population 4,652 randomised to Dec 2018 (3% (142), 6% (254), 5% (242) and 0.4% (20)) and as these require separate applications to registry data these were not accessed within the timelines for this feasibility project. In India there is not an equivalent database to gain this data from.

Data from relevant CRFs that had been received and entered on to the trial database by the MRC CTU in April 2020 was extracted by Fay Cafferty (FC), Add-Aspirin trials unit lead and senior statistician, and uploaded into the UCL data safe haven. Data received after this extraction date was not compared. A summary of data extracted included:

1. Registration and randomisation date
2. Comorbidity data at randomisation
3. Details of initial cancer diagnosis (diagnosis date and staging) and treatment prior to randomisation for all four cohorts
4. Cancer recurrence data (type of recurrence and timing) and new primary cancers
5. SAE data defined as per protocol and CRF follow-up questions (date and type of SAE)
6. Death data (date and cause)

3.2.3.2 NCRAS data

NCRAS has been described within the introduction of this thesis and therefore detail of this registry will not be discussed extensively in this chapter. Data was collected from different datasets either held by NCRAS or that could be linked via the NCRAS service. Depending on the source and the nature of the data there were varying censor

dates that needed to be considered for the analysis. For each linked participant, all tumours (C00-D48) diagnosed from 01 December 2012 to 30 June 2018 were provided from cancer registration data. This was the most up to date registration data at the time of data access. Cancer registration data provided detail on ICD-10 code of cancer diagnosis, date of diagnosis, particular histological staging detail, and predictive/prognostic markers depending on the tumour type. The exact data requested from PHE is attached in **Appendix C**.

HES data held by NHS Digital but linked by NCRAS include hospital inpatient, outpatient and A&E data. HES episodes were limited to records from 31 days prior to diagnosis of the earliest first primary of the solid tumours within the study (ICD-10 codes C50, C18-C20, C15, C16, C61) up to 31 March 2019 (most recent available data at the time of data release). The NHS Digital CWT dataset which provides detail on diagnosis (primary tumour/local or distant recurrence), all cancer treatments and intent of treatment also had a censorship date of 31 March 2019. ONS mortality data was provided by NCRAS up to December 2019. DIDs provided information about imaging performed for each participant in their treatment pathway up to March 2019 but was not included in this feasibility analysis.

NCRAS datasets including SACT dataset and RTDS were also censored in March 2019. These two datasets gave information on timing and treatment detail for all patients who have had systemic anti-cancer treatment or radiotherapy.

3.2.4 Statistical Analysis

Analyses were planned to be largely descriptive (since numbers of events for each outcome were anticipated to be relatively small) and included:

- Number (%) of trial participants that it was not possible to identify in the routine datasets (overall and by tumour specific cohort) using a consort diagram. Numbers for consort diagram provided by FC for trial data and NCRAS analysts for registry data.

- Details of cancer registration data in the trial compared to NCRAS data using Cohen's (weighted) Kappa coefficient, with 95% confidence interval, to assess concordance of categorical data between two "raters" (or in this case databases) to what might be expected by chance. Results were between 0 – 1, with a number close to 1 indicating good agreement. This is presented by tumour group.
- Details of primary outcome events in the trial that it was not possible to obtain information on from routine datasets. Data was only compared up to the censorship dates in the NCRAS data and any trials events that lay outside this were excluded.
- For each outcome event where it was possible to obtain routine data, concordance between events identified in the trial database and those identified in routine datasets were presented with descriptive statistics, tabulations and percentages. For disease recurrence, this was repeated for each tumour-specific cohort.
- Sites were contacted to confirm death and recurrence data where there was EHR data indicating an event but no equivalent trial data.
- All statistical analyses were performed using STATA (version 16) depending on analysis by AM or FC (see below)

3.2.5 Comparison analysis methods and definitions

3.2.5.1 Trial baseline cancer registration

Cancer registration data was compared with trial cancer registration data at time of trial enrolment for randomised participants. For each tumour type the TNM staging for malignant tumours was assessed for conformity between the two datasets. As multiple tumours can be registered for a patient this was assessed against the appropriate ICD-10 code for the trial participant and date of diagnosis. Firstly the number of trial participants with no cancer registration staging information was recorded and the comparison was only performed between those with data in both data sets.

NCRAS have two TNM staging categories. These are histological staging (Tpath, Npath and Mpath) and also best staging (Tbest, Nbest and Mbest). Best staging is considered the most appropriate staging created by trained cancer registration officers from multiple different sources including radiological, histology and MDT documentation. The comparison was performed between histological staging and Add-Aspirin histological staging where possible following surgery. Best staging was also assessed versus Add-Aspirin trial staging that would be used for analysis either post-surgical histological staging or pre-treatment staging where histological staging was absent. Neo-adjuvant treatment before surgery could change the staging and affect the comparison and therefore this was a separate variable to consider. Each tumour type has different treatment pathways and also subtle differences in TNM staging. Therefore, the analysis was based on prior treatment and the TNM staging for each tumour. This was particularly important where there was neoadjuvant treatment and also radical radiotherapy instead of surgery. **Table 3.1** shows additional tumour specific staging variables that were compared. The colorectal cancer cohort had a small number with liver metastasis which were resected on a curative pathway and the metastatic staging was reviewed in this cohort. This analysis was designed and carried out by myself and was checked by FC. Kappa coefficient analysis was also carried out by FC.

Table 3-1: Tumour specific staging comparisons

Tumour Cohort	Staging comparison parameter
Breast	Oestrogen receptor (ER) status
Colorectal	Metastatic staging
Prostate	Gleason Score

3.2.5.2 Death data comparison

A death data comparison was analysed as per a Study within a trial (SWAT) protocol written within the MRC CTU (in which I was a collaborator). This protocol was placed in the Northern Ireland Hub for Trials Methodology Research SWAT repository store (SWAT 125: Comparison of trial-collected and routinely-collected death data) (160). This analysis was carried out by myself using STATA coding and checked by FC.

A comparison of all deaths within the trial and mortality data within NCRAS data was performed up to 31st December 2019 due to censorship date of NCRAS data. Descriptive statistics were used to compare percentage difference, sensitivity and PPV for ascertaining where a death had occurred using trial data as the 'gold standard'. Cohen's Kappa coefficient was also included in the analysis. Comparisons were performed for the following parameters:

1. Vital Status: death recorded Yes/ No
2. Death date and if there is a discrepancy in date then the median difference between the two datasets and range is described
3. Death cause

3.2.5.3 Serious Adverse Events (SAE)

Specific SAEs as defined by the protocol as secondary outcomes were compared with trial data either through follow-up forms or through SAE data submitted by sites following trial registration. Events were compared with HES APC (inpatient) data up to a censor date of 31st March 2019. Sets of ICD-10 codes for identifying the events were pre-defined prior to the analysis. For acute events as defined in the protocol (gastrointestinal (GI) haemorrhage, intracranial haemorrhage, myocardial infarction, cerebral infarction and thrombotic events) these were compared with the first event of each type in the HES data. Only diagnosis in box 1 of the HES dataset was used which prevented events being counted more than once due to prolonged or multiple admissions. Date of event was also compared with the trial data.

Chronic conditions (diabetes, dementia and macular degeneration) were also compared against the first entry in HES data but all diagnosis boxes were used as

they may not be necessarily be admitted for the condition coded. All participants with a diagnosis of these conditions prior to randomisation were excluded from the analysis. This analysis was designed by myself and carried out using STATA coding by FC.

ICD-10 codes included in **Table 3.2**:

Table 3-2: Adverse events as per protocol and ICD-10 codes used

Adverse event as per protocol	ICD-10 Codes used
Serious Haemorrhage	Limited this to upper/lower GI haemorrhage and intracranial bleed as per previous publications. Due to complexity of coding, previous publications codes were used and are presented in **Appendix D**. (161)
Serious vascular events: Acute myocardial infarction and cerebral infarction	I21 Acute myocardial infarction I22 Subsequent myocardial infarction I23 Current complications following acute myocardial infarction I24.9 Acute ischaemic heart disease, unspecified I63 Cerebral Infarction I64 Stroke not specified as haemorrhage or infarction (thromboembolic stroke) G45 Transient cerebral ischaemic attacks and related syndromes
Thrombotic events	I26 Pulmonary Embolism I80.2 Phlebitis and thrombophlebitis of other deep vessels of lower extremities including deep vein thrombosis not otherwise specified (NOS) 180.3 Phlebitis and thrombophlebitis of lower extremities, unspecified including embolism or thrombosis of lower extremity NOS I81 Portal vein thrombosis I82 Other venous embolism and thrombosis

Diabetes and associated complications	E10 Insulin dependent diabetes mellitus E11 Non insulin dependent diabetes mellitus E13 Other specified diabetes mellitus E14 Unspecified diabetes mellitus
Diagnosis of dementia	F00 Dementia in Alzheimer disease F01 Vascular Dementia F02 Dementia in other diseases classified elsewhere F03 Unspecified dementia
Macular degeneration	H353 Degeneration of macula and posterior pole

3.2.5.4 Other new primary cancer

Other new primary cancers were reviewed using cancer registration data and CWT data. This had a cut-off point of 31st March 2019 with the recognition that cancer registration data is censored to 30th June 2018 so there was nine months where only CWT data was reviewed. This was compared with trial records for new primary cancer. Comparisons were described using tabulations due to small numbers. This analysis was done by myself using STATA coding and manual review.

3.2.5.5 Cancer recurrence

Initially cancer recurrence was reviewed using cancer registration data and CWT data. This had a cut-off point of 31st March 2019 with the recognition that cancer registration data is censored to 30th June 2018 so there was nine months where CWT data was only reviewed. The recurrence was defined as any documentation of recurrence as per predefined codes as per **Table 3.3** following randomisation. This was compared with trial recurrence data to see if date and recurrence type is comparable. Prostate specific antigen (PSA) progression for prostate cancer was not assessed in the recurrence comparison as, after discussion with NCRAS analysts, this was deemed unfeasible. This analysis was done by myself using STATA coding and manual review of results.

Table 3-3: Definition of codes used for recurrence in cancer registration and cancer waiting times dataset

Cancer waiting time/ cancer registry variable	Code required
Diagnosis ICD-10 code	C77 (Secondary and unspecified malignant neoplasm of lymph node) C78 (Secondary malignant neoplasm of respiratory and digestive organs) C79 (Secondary malignant neoplasm of other and unspecified sites) (within the cancer waiting times dataset or Original ICD-10 code following randomisation in cancer registration dataset)
Patient status and NHS treatment status	15-20 (Cancer recurrence suspected or diagnosis codes) 22 (Cancer recurrence code) 30-36 (Cancer progression code)
Cancer Treatment Event type (Reason for treatment)	3-12 (Treatments for recurrence, relapse, transformation of progression, or metastatic disease)
Treatment modality	07 (Specialist palliative care)
Metastatic site code	Any code in variable

3.2.5.6 Cancer recurrence review of known recurrence in NCRAS data

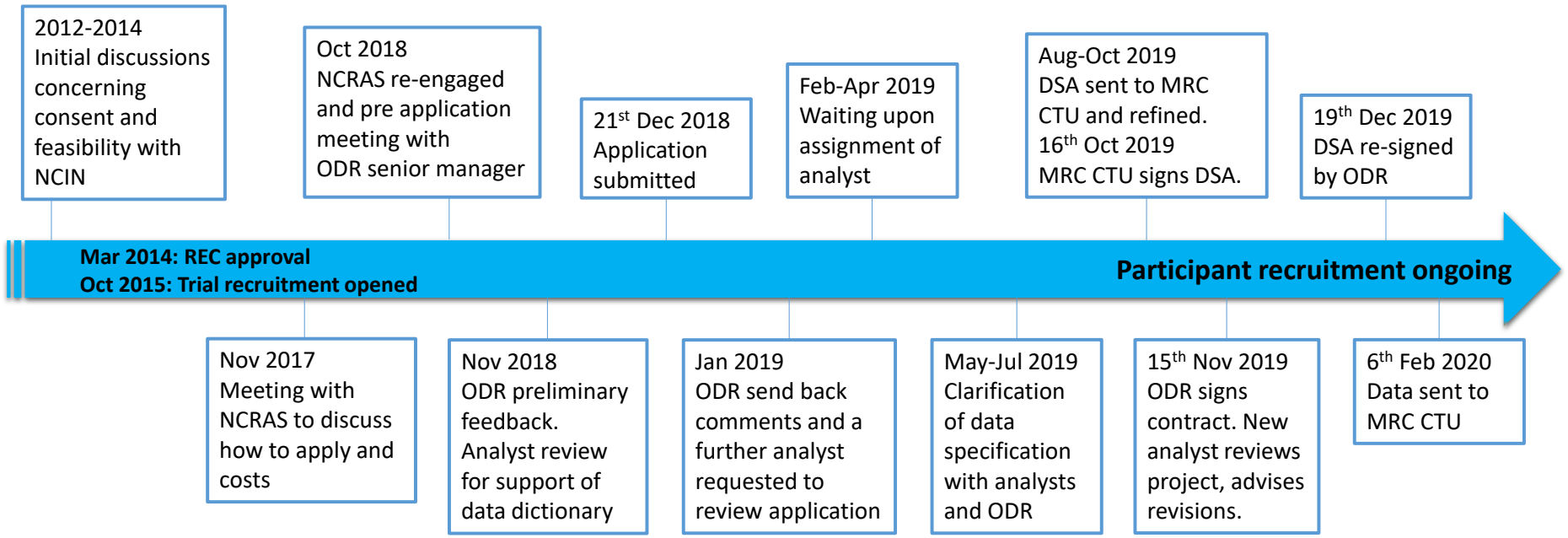
All trial recurrence events with no corresponding event in the cancer registry/CWT data base were reviewed. For these events the HES APC data using recurrence codes C77-C79, SACT/ RTDS for new systemic treatments or newly prescribed radiotherapy was also examined. All three datasets only used data following randomisation to establish if this was a recurrence event based on the premise that all patients should be cancer free at the time of randomisation and not be starting new treatments or classified as having metastatic disease. This analysis was done by myself using STATA coding and manual check of the data. Recurrence algorithms could also be used in this setting but this is out of the remit of this project.

3.3 Results

3.3.1 Data access and completeness of participant data

The data request to ODR for data held by NCRAS was submitted December 2018, and data was made available early 2020. The time course and necessary requirements for the data release has been documented in **Appendix B** and published (120). The flow chart of data access is summarised in **Figure 3.3**. Trial Data was frozen April 2020 and extracted to be evaluated with NCRAS data.

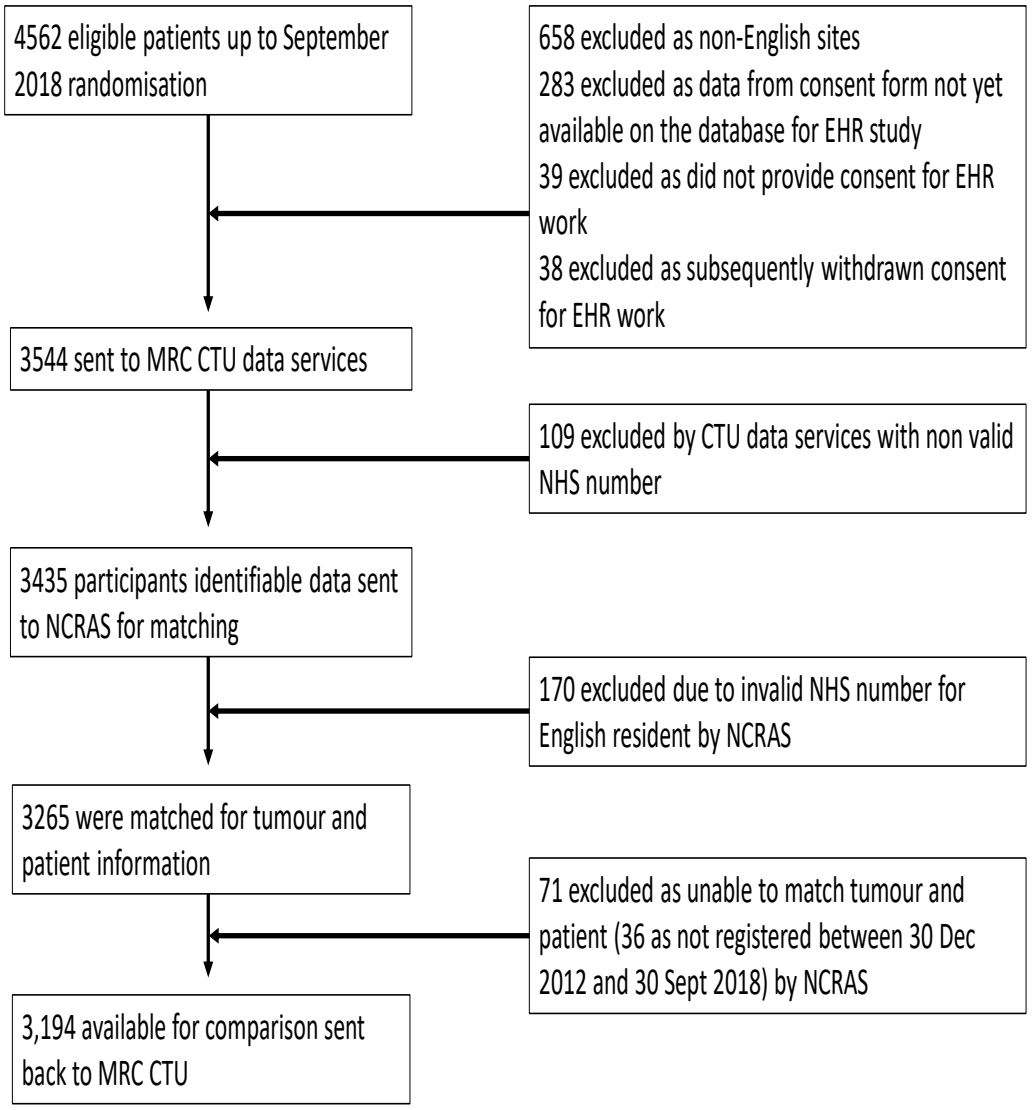
Figure 3-3: Flow diagram of the Add-Aspirin National Cancer Registration and Analysis Service (NCRAS) application. (Please note that timeline is not proportional) Adapted for the purposes of this thesis from Macnair Trials 2021 publication. (120)



CTU: Clinical Trials Unit; DSA: Data Sharing Agreement; NCIN: National Cancer Intelligence Network; ODR: Office for Data Release; REC: Research Ethics Committee

In total 4,562 participants were potentially available for this analysis (defined as having been randomised into the Add-Aspirin trial between 1st December 2015 and 30th September 2018). Patients were excluded from the analysis (and transfer of their details to NCRAS) either because the treating site location was outside of England, invalid NHS numbers or declined consent to use of their routine data. This meant that 3,435 participants' identifiable data were sent to NCRAS for linkage **Figure 3.4**. NCRAS provided information on why they were unable to link some of the patients to their data sets as described in **Figure 3.4**.

Figure 3-4: Consort diagram of eligible participants for comparison



Overall 3,194 out of the available 3,544 (90%) eligible for the project were identified for comparison. 109 (3%) excluded by MRC CTU data services and a further 241 (7%) were excluded following NCRAS review. The final cohort for analysis was 3,188 due to 6 further participants being excluded at time of analysis as they had retracted their consent for data to be used for this project between the time of submitting the data request to NCRAS and the time of analysis.

On review of the four cohorts the percentage that could be identified within NCRAS were breast (1489 (92%)), colorectal (866 (90%)) and prostate (704 (87%) cohorts was similar as seen in **Table 3.4**. For the gastro-oesophageal cohort (129 (91%)) were identified noting that the number of patients recruited overall is lower in England and a higher percentage within the trial are from India.

The total cost of the one-off data extraction was £6,048. This was decided by ODR/ NCRAS based on the complexity of the data extraction and number of extractions. The guide price that was published by the ODR was based on an hourly rate and this was dependent on the request, so pricing is bespoke to the application. On planning for future extractions ODR stated that the cost would be approximately £415.80 for an amendment to the original contract and then initially approximately £831.60 per annual extraction. This was based on the premise that we would extract the same data and that the cohort would increase in number due to continued recruitment. Once the trial had completed recruitment this would decrease further to £415.80 based on current rates defined by the ODR. (162)

Table 3-4: Number of available participant data per tumour group

Tumour Type	Number potentially available for analysis prior to exclusion by MRC CTU data services and NCRAS	Numbers within analysis once EHR data provided
Breast	1627	1489 (92%)
Colorectal	960	866 (90%)
Gastro-oesophageal	142	129 (91%)
Prostate	809	704 (87%)

3.3.2 Cancer registration comparison

3.3.2.1 Breast

1,489 (92%) out of 1,627 patients had registry records to compare with trial data (**Table 3.5)**. Data was missing for some of the variables within registry staging depending on the variable. It was noted that there was a relative low level of missing data within best T staging (Tbest) staging (1,444 (97%)) compared with T histology (Tpath) (1,083 (73%)). However, the concordance between post-surgery trial staging was better with Tpath (95%, Kappa Coefficient (κ)= 0.76) than with Tbest (92%, κ=0.65). This was also the case with N staging (Npath 94%, κ=0.81 and Nbest 92%, κ=0.73). When participants who had neo-adjuvant treatments were removed the rate of conformity increases for Tbest (94%, κ=0.75) and Nbest (94%, κ=0.80). This demonstrates the possible subjectivity of Tbest in the documentation of pre or post neo-adjuvant treatment staging. Tpath (96%, κ=0.77) and Npath (94%, κ=0.81) in this situation does not change significantly as this is based on post-surgery and therefore post neo-adjuvant treatment.

Table 3-5: Concordance of registry staging and trial staging for breast cohort

Registry data category	All breast participants post- surgery trial staging cohort (n=)	All breast participants post- surgery trial staging observed agreement	Cohen's (weighted) Kappa coefficient with 95% confidence interval	Without Neoadjuvant patients only post-surgery trial staging cohort (n=)	Without Neoadjuvant patients only post-surgery trial staging observed agreement	Cohen's (weighted) Kappa coefficient with 95% confidence interval
Tbest	1444	91.7%	0.65 (CI 0.61-0.69)	1042	94.0%	0.75 (CI 0.70-0.95)
Tpath	1083	95.2%	0.76 (CI 0.69-0.78)	794	95.9%	0.77 (CI 0.72-0.82)
Nbest	1045	91.5%	0.73 (CI 0.70-0.77)	1041	93.6%	0.80 (CI 0.76-0.85)
Npath	1072	94.0%	0.81 (CI 0.77-0.86)	790	93.9%	0.81 (CI 0.76-0.86)

Table 3-6: ER status concordance for breast cohort between registry and trial data

ER status (+/-)	ER status (+/-) cohort (n=)	Percentage concordance	PPV	Sensitivity
ER	1156	97%	99%	97%

On review of tumour specific staging details, with respect to ER status (**Table 3.6**), a positive or negative ER result was documented in PHE data in the majority of participants 1,156 (78%) The concordance between those with a result and trial data is 97% with PPV 99% and sensitivity 97%.

3.3.2.2 Colorectal

866 (90%) out of 960 participants had registry records in the colorectal cohort for comparison with trial data (**Table 3.7**). Only 19 patients had neo-adjuvant chemotherapy a separate analysis was not done for this cohort. The 99 participants that received radiotherapy prior to surgery were analysed separately as this may significantly affect staging. Colorectal patients who had had liver metastases resected are eligible for the trial therefore M stage was also assessed in the TNM staging.

When patients without radiotherapy were analysed, there was good correlation in terms of staging between trial data and that derived from NCRAS. Specifically, Tbest (99%, κ=0.90) and Tpath (99%, κ=0.93), Nbest (98%, κ=0.94) and Npath (99%, κ=0.97) and Mbest (98%, κ=0.69) all had very good concordance. The results of this were very similar to the total cohort due to very similar number of participants. Interestingly this dropped when patients who had radiotherapy were analysed for Tbest (89%, κ=0.30), Nbest (68%, κ=0.27). However, this did not seem to be an issue when Tpath (97%, κ= 0.86) and Npath (98%, κ=0.93) and Mbest (99%, κ=0.79) were reviewed. Tbest (94%, κ=0.58) did improve when this was compared with pre-treatment trial data. It is noted again that there was a low level of missing data within Tbest (846 (98%)) and similar level of missing data for Tpath (836 (96%)) as well.

Table 3-7: Registry data and trial data concordance for colorectal cohort

Registry data category	All colorectal participants post-surgery trial staging cohort (n=)	All colorectal participants post-surgery trial staging observed agreement	Cohen's (weighted) Kappa coefficient with 95% confidence interval	All colorectal participants without radiotherapy post-surgery trial staging cohort (n=)	All colorectal participants without radiotherapy post-surgery trial staging observed agreement	Cohen's (weighted) Kappa coefficient with 95% confidence interval	All colorectal participants with radiotherapy post-surgery trial staging cohort (n=)	All colorectal participants with radiotherapy post-surgery trial staging observed agreement	Cohen's (weighted) Kappa coefficient with 95% confidence interval
Tbest	849	97.6%	0.84 (CI 0.79-0.89)	756	98.5%	0.90 (CI 0.85-0.95)	85	89.1%	0.30 (CI 0.18-0.41)
Tpath	836	98.8%	0.93 (CI 0.88-0.97)	745	99.0%	0.93 (CI 0.88-0.99)	83	97.3%	0.86 (CI 0.72-1.00)
Nbest	742	94.5%	0.86 (CI 0.81-0.91)	755	97.6%	0.94 (CI 0.88-0.99)	87	67.8%	0.27 (CI 0.13- 0.41)
Npath	831	98.6%	0.96 (CI 0.91-1.00)	741	98.7%	0.97 (CI 0.91-1.00)	82	97.6%	0.93 (CI 0.76-1.10)
Mbest	742	98.4%	0.69 (CI 0.62-0.76)	742	98.4%	0.69 (CI 0.62-0.76)	87	98.9%	0.79 (CI 0.59- 1.00)

3.3.2.3 Prostate

704 out of 809 (87%) patients had registry records that could be compared to trial staging data (**Table 3.8**). Prostate cancer participants receive either surgery or radical radiotherapy as the definitive treatment. Participants who had radiotherapy would not have post-surgery histology (Tpath/ Npath) and therefore Tbest/Nbest was compared with trial clinical staging. These two radical treatment groups were analysed separately. T staging is also broken down to more accurate staging in both datasets (e.g T2a, T2b) rather than broad groups. Comparisons was done on broad categories of T1, T2, T3. It is noted that again there was a low level of missing data with Tbest (689 (98%)). Missing Tpath was not assessed as there was a large amount of radiotherapy patients who would not have had surgery.

In the surgical group Tpath (99%, κ=0.94) had a significantly higher concordance rate then Tbest (90%, κ=0.64). In the radiotherapy group Tbest (98%, κ=0.69) which showed a slightly better concordance then Tbest in surgical patients. Nstaging had an extremely good concordance of near 100% in both the surgical and radiotherapy group. Cohen's (weighted) Kappa coefficient was not assessed in this group due to very high concordance rate and that it was comparison between N1 and N0.

Gleason score was also compared with trial data (**Table 3.9**). This demonstrated a good concordance of Gleason score (96%, κ=0.81) between trial and registry data. There was also a low level of missing data with 687 (98%) having Gleason scores to compare between registry and trial data.

Table 3-8: Staging comparison between registry data and trial data for prostate cohort

Registry data category	All prostatectomy prostate participants post-surgery trial staging cohort (n=)	All prostatectomy prostate participants post-surgery trial staging observed agreement	Cohen's (weighted) Kappa coefficient with 95% confidence interval	All radiotherapy prostate participants with clinical trial staging cohort (n=)	All radiotherapy prostate participants with clinical trial staging observed agreement	Cohen's (weighted) Kappa coefficient with 95% confidence interval
Tbest	296	89.7%	0.64 (CI 0.54-0.74)	393	98.1%	0.69 (CI 0.61-0.76)
Tpath	280	98.6%	0.94 (CI 0.83-1.00)	N/A	N/A	N/A
Nbest	275	99.6%	0.80 (CI 0.68-0.91)	356	100%	N/A
Npath	114	100%	N/A	N/A	N/A	N/A

Table 3-9: Gleason comparison between registry data and trial data for prostate cohort

	Gleason score cohort (n=)	Gleason score percentage observed agreement	Cohen's (weighted) Kappa coefficient with 95% confidence interval
Gleason Score in trial data and registry data	687	95.5%	0.81 (CI 0.76- 0.87)

3.3.2.4 Gastro-oesophageal

129 (91%) out of 142 participants had registry records that could be compared with trial data (**Table 3.10**). It again noted that there was a low level of missing data in Tbest (119 (92%)) which was better when compared to Tpath (100 (78%)). The Gastro-oesophageal cohort was not broken down by treatment as the cohort was much smaller than the other tumour groups. Also the treatment was variable with either chemoradiotherapy or surgery. The treatment pathway is further complicated with participants possibly receiving neoadjuvant and adjuvant chemotherapy as well. Therefore with small numbers and variable treatment regimens the comparison was with the whole cohort.

The gastro-oesophageal cohort had the worst concordance of T best (84%, κ=0.32) and N best (80%, κ=0.41) which probably reflects the complexities of treatments. The concordance greatly improves when Tpath (94%, κ=0.81) and Npath (95%, κ=0.88) are compared to registry data.

Table 3-10: Staging comparison between registry and trial data for gastro-oesophageal cohort

Registry data category	All gastro-oesophageal participants post-surgery trial staging cohort (n=)	All gastro-oesophageal participants post-surgery trial staging observed agreement	Cohen's (weighted) Kappa coefficient with 95% confidence interval
Tbest	119	84.5%	0.32 (CI 0.21-0.43)
Tpath	100	94.3%	0.81 (CI 0.68-0.94)
Nbest	118	80.0%	0.41 (CI 0.29- 0.54)
Npath	98	95.2%	0.88 (CI 0.73- 1.00)

3.3.3 Death data comparison

Death was recorded by ONS but provided by NCRAS service which was censored at 31st December 2019. Trial data was frozen and extracted in April 2020 but censored at 31st December 2019 for this comparison. There were 134 deaths reported in the trial data during the period of interest. 24 of the 3,188 participants known to have data from the registry had no death data as they had a no vital status information (blank cell/ not followed up). **Table 3.11** displays the comparison between the two datasets.

Death data had good concordance between trial and registry data (κ=0.91). If trial data is stated to be the gold standard then the PPV of the registry data is 85% and sensitivity of 100%. This may be misleading in this case as what this demonstrates that the registry data is extremely good at picking up deaths. The PPV in this case demonstrates probable true deaths. Also, out the 23 registry deaths that had no trial data 1 had been lost to follow-up or 2 had withdrawn from follow-up. 10 had a prolonged time without follow-up prior to the registry death (which may indicate that they had been lost to follow-up, though this had not been formally documented). Sites were contacted about the 20 patients who were identified as having died through the registry data and were still considered in active follow-up with no death recorded at site. 9 out of the 20 registry deaths, during a period of 6 months, still had no confirmation of death from sites. The remaining 11 deaths were confirmed by site with subsequent CRF submission.

When the date of death was compared 9 out of the 134 (6.7%) were different with the median discrepancy of 3 days and a range of 1-365 days. There were two outliers of 91 days and 365 days and were subsequently verified with the site showing that EHR data was correct. Lastly the cause of death was reviewed and there was a discrepancy in 15 out of 114 cases (10.4%). The difference was mainly due to a missing cause of death in either the registry or trial data (12 out of 15) which were balanced between the two. Only 3/15 had differences in documented cause of death when data was entered in both registries.

Table 3-11: Trial vital status versus registry data vital status

Vital status registry data	Trial data			Cohen ‘s Kappa Coefficient with 95% confidence intervals
	Death	Alive	Total	
Dead	134	23	157	0.91 (CI 0.88-0.95)
Alive	0	3007	3007	
Total	134	3030	3164	

3.3.4 Adverse event comparison

3.3.4.1 Serious GI and intracranial bleeds

Within the trial data, 11 serious bleeds (GI or intracranial, CTCAE grade 3 or higher) had been reported in 11 patients up to the end of March 2019. Five of these events could also be identified in the HES APC dataset, with diagnosis dates all within 4 days of the trial report. The remaining six patients (all of whom had lower gastrointestinal, CTC grade 3 bleeds reported in the trial dataset) did not have GI bleeding events reported in the HES dataset (considering only diagnosis box 1).

There were also 40 GI or intracranial bleeding events in HES APC (diagnosis box 1), for 39 patients, which are not in the trial dataset. This includes two patients who did have a matched event (reported both in the trial and HES), but additionally had an earlier bleed reported in HES.

16 patients with a bleed identified in HES, there is some indication of a bleeding event within the trial data, but it has either been reviewed as not meeting the serious bleed criteria or it has not yet been reviewed (in most cases this is because it has been reported but a full report had not yet been received). This demonstrates that the bleeding event in HES may not be significant enough to be defined as a trial serious event.

For the remaining 21 patients with a bleed identified in HES, we have no record of a bleeding event within the trial; yet all of these patients were still being followed up in the trial at the time of the event and 14 have follow-up data beyond the HES event date. These were mostly GI unspecified bleeding events (ICD 10 K922) in HES (12/21).

3.3.4.2 Vascular events

Within the trial cohort, there were 14 vascular events (in 14 patients) reported before the end of March 2019, including MI, cerebral infarction or Transient Ischaemic Attack. Six (43%) of these events could also be identified in the HES APC dataset, with

diagnosis dates all within 3 days of the trial report. 8 of the events were not found in HES.

There were five vascular events identified in HES for patients with no event reported within the trial. These had a range of diagnosis codes (TIA, 2 cerebral infarcts and 2 MIs). One of these patients had withdrawn from trial follow-up prior to the event; the other four remained in follow-up and all but one had trial data from after the date of the HES event.

3.3.4.3 Thrombotic events

Within the trial cohort, there were 24 thrombotic events (in 24 patients) reported before the end of March 2019, including PE and DVT. Five (20.8%) of these events were identified within the HES APC dataset with diagnosis dates within 4 days of the trial report. The remaining 19 events could not be identified within HES APC; these include both PE, DVT and other events occurring over a range of dates.

There were also 14 events (in 13 patients) identified in HES APC which had not been reported in the trial; most of these were PEs (ICD 10 I269). Whilst all of these patients remained in trial follow-up at the time of the event, only three had follow-up data from after the HES event date.

3.3.4.4 Diabetes

Within the trial cohort, 16 new diagnoses of diabetes or diabetic complications were reported before the end of March 2019. Three of these were complications (retinopathy) in patients reported to have diabetes at baseline; the rest were new diagnoses of diabetes in patients without the condition at baseline.

4 of the 16 (25%) patients were found to have diabetes/a diabetic complication reported in the HES dataset, and the first report was between 0 and 5 months after the trial report. The remaining 12 patients had no indication of diabetes or diabetic complication within HES APC.

Diabetes or a diabetic complication were additionally identified in HES for 22 (of 3216) patients who did not have diabetes reported at registration and did not have a diagnosis reported during follow-up in the trial. These were mostly reports of type 2 diabetes without complications (ICD E119). All of these patients were still being followed up in the trial at the time of the HES report and all but three had trial data from after this date.

3.3.4.5 Dementia

Within the trial cohort, three new diagnoses of dementia (in patients without the condition at the time of registration) were reported before the end of March 2019. Only one of these patients (33%) was found to have dementia reported within the HES dataset, approximately 17 months after the trial report. The remaining two patients had no indication of dementia within HES APC. It is notable that, for one of these, the diagnosis date reported within the trial was February 2019, close to the most recent date within the HES dataset. Dementia was, however, reported for one further trial patient who did not have the condition reported either at registration or during follow-up within the trial. This patient was still being followed up within the trial and had follow-up data for more than a year beyond the earliest report of dementia within HES.

3.3.4.6 Macular degeneration

Within the trial cohort, five new diagnoses of macular degeneration were reported before the end of March 2019 (note that, as this was an exclusion criterion, no participants had the condition at the time of registration). Only one of these patients (20%) was found to have macular degeneration reported within the HES dataset, approximately 7 months after the trial report. The remaining four patients had no indication of macular degeneration within HES APC. Macular degeneration was, however, reported for eight further trial patients who did not have the condition reported during follow-up within the trial. These patients were all still being followed up within the trial at the time of the HES report and all but two had trial data from after this date.

3.3.5 Second primary cancer

CWT data and cancer registry data was compared with trial data with the knowledge of differing censorship dates between the two datasets of March 2019 for CWT and June 2018 for cancer registration data. A second primary was defined as a registered cancer not in keeping with the cancer registered for the trial either with different ICD-10 code or a second primary with the same ICD-10 code in the cancer registry. Participants could have a diagnosis of skin cancer before trial registration so any diagnosis of a new skin cancer post registration could be a recurrence which would not be known about without verifying with the site. Also not all skin cancers were defined as a second malignancy as per the protocol like basal cell carcinomas. Therefore, any skin cancers that are found to be in the registry data may not have trial data due to the above reasons and would not be defined as missing in the trial data.

In the breast cohort, there were 5 non breast primaries recorded in trial data before March 2019 and all of these were recorded in the registry data. There was a total of 3 new primary cancers (non second breast cancers and non skin cancers) which were found in the registry data and not the trial data. All were in follow-up at the time of analysis. There were 4 new contralateral breast primary cancers within the time period and 3 were picked up by both trial and registry data and 1 where the registry data had no record.

In the colorectal cohort there were 13 events defined as a new primary cancer events within the trial prior to March 2019. 2 were not found in the registry data. 4 further new primaries were found in the registry data and not in the trial data which were not skin cancers. All were in follow-up at the time of analysis. No second colorectal primaries were documented in trial data prior to March 2019 that were not also defined as loco regional or distant metastasis.

In the prostate cancer cohort there were 20 new primary cancer events prior to March 2019 in the trial dataset. 8 had no registry data for the new primary cancer event. 2 of these had either a recurrence event or a death date at a very similar date and 3 were very close to the registry censor date. There were 2 new primary cancer events that

were found within the registry data that were not in the trial data that were not skin cancers. All of these were in active follow-up.

In the gastro-oesophageal cohort there were 2 new primary cancer events in the trial data prior to March 2019. Both of these were in the trial data and registry data. There were no new primary cancers in the registry data and not the trial data which were not a skin cancer.

3.3.6 Recurrence

The initial review of recurrence was achieved primarily through the CWT dataset. This was due to the fact that this was the only dataset that had official codes for recurrence or relapse and defined them as local or metastatic. This was achieved by looking for a defined set of recurrence codes following registration in each cohort separately. This was censored at 31 March 2019.

In the breast cohort 39 (45%) of the total 87 recurrences in the trial dataset had CWT data that demonstrated recurrence following randomisation. In the colorectal cohort 46 (42%) of 110 recurrence events in the trial dataset had CWT data that demonstrated recurrence following randomisation. In prostate cancer cohort 13 (35%) out of the total 37 recurrences in the trial dataset had CWT data that demonstrated recurrence following review. Lastly in the gastro-oesophageal cohort 17 (44%) of 39 recurrences in the trial dataset had CWT data that demonstrated recurrence following review (**Table 3.12**). Many of the events that were missed as CWT had no data following the randomisation date with 63 (23%) out of 273 total events in all cohorts. When recurrence data was matched between the trial and registry data the majority had a similar date of recurrence and nearly all matched if metastatic or local recurrence.

Table 3-12: Percentage of recurrence captured with CWT dataset using predefined codes and percentage recurrence using exploratory analysis of all datasets

Tumour group	Number of recurrences prior to March 2019 in trial data	Percentage of trial recurrences captured in CWT dataset only	Percentage of trial recurrences captured when all datasets used	Percentage of trial recurrences captured when all datasets used matched to within 1 month
Breast	87	39 (45%)	82 (94%)	79 (91%)
Colorectal	110	46 (42%)	104 (95%)	78 (71%)
Prostate	37	13 (35%)	30 (81%)	28 (76%)
Gastro-oesophageal	39	17 (44%)	35 (90%)	33 (85%)

Table 3-13: Number of potential recurrence events captured by registry data which was not in trial data and percentage confirmed true events

		Number of additional potential events in CWT alone			
Tumour type	Number of recurrences prior to March 2019 in trial data	Total	Number lost to follow-up/ no confirmation response form site	Number verified as a true recurrence	Number with reported as pre-malignant disease or no recurrence
Breast	87	11	4	5	2*
Colorectal	110	8	3	5	0
Prostate	37	2	0	0	2[$]
Gastro-oesophageal	39	0	N/A	N/A	N/A

* DCIS or persistent disease [$] Site confirmed no recurrence

The review of CWT data using defined codes also demonstrated that there were 21 potential recurrence events in the registry data that had not been picked up in the trial data (**Table 3.12**). When these events were reviewed 2/21 patients had been lost to follow-up. When we contacted the site to verify the recurrence we were unable to get response from the site for 5 of the events. 10 were verified as true events and subsequent CRFs were sent in. 4 were deemed to not be true events as one was persistent disease, one was ductal carcinoma in situ (DCIS) and 2 prostate recurrence were considered not to be true events by the site. (**Table 3.13**).

Following the review in **Table 3.12** further work was done with the cancer registration dataset, CWT, SACT, RTDS and HES APC datasets to see if a simple algorithm could identify all the recurrences using all the datasets. This was possible as the trial is an adjuvant study and has a clear time period post randomisation where participants should be cancer and cancer treatment free. All participants, with a known recurrence, were reviewed to see if that had a new surgical, radiotherapy or anti-cancer treatment code following randomisation in CWT, SACT and the radiotherapy dataset. HES APC data was reviewed to see if a new ICD-10 code for metastatic disease was noted following randomisation. These measures meant that nearly all events (summarised in **Table 3.12**) could have been picked up using systemic codes that could be refined and put in an algorithm in future work to pick up recurrences from EHR data. It differed by tumour type which dataset and event that picked up the recurrence and there was a delay in the recurrence compared to the trial data as it was based on a treatment rather than diagnosis. This was exploratory analysis based on clinical review. In breast/ colorectal and gastro-oesophageal cancer SACT and HES APC data were the best datasets to confirm recurrence due to participants starting new chemotherapy regimens or an inpatient admission which was coded as metastatic disease ICD-10 code. In prostate cancer due to the recurrence following surgery this was best confirmed with the radiotherapy dataset with salvage radiotherapy.

3.4 Discussion

This study has reviewed different aspects of trial outcomes and also staging information within English NCRAS registry data versus trial data. Both outcomes and

staging are important to evaluate the use of EHR records from NCRAS for the Add-Aspirin long-term follow-up strategy and also in the future more broadly for oncology trials. We have demonstrated the variability of datasets compared to trial outcome data.

Registry mortality data demonstrated 100% sensitivity with trial data and also demonstrated deaths that were found to be not yet reported in the trial data. The trial date of death and also cause of death had a high degree of concordance with registry data. This confirms that death registry data is a valid resource to confirm deaths in the trial. In general it had a smaller lag time then trial data and less susceptible to loss to follow-up. This also supports previous literature in this subject on the validity of using mortality registry data (60, 144-148, 152). Mortality data maybe the most important registry to have for a trial and evidence is mounting that it could possibly replace trial mortality data. However, it may be more successful to not only confirm trial deaths at the end of the trial but also to help with interim analysis while trials recruit to help IDMCs to assess survival data and quality of trial data collection.

Recurrence data initially looked like there was poor comparability between the trial data and registry data with under half of the recurrences being picked up by the CWT registry which would be consistent with previous work on the topic (75, 153, 155). It did demonstrate that if CWT was used then the type of recurrence (local or metastatic) was consistent with the trial data and had a similar date of recurrence. The trial had an advantage in that it had a defined date where the patient should be disease free and should have no new anti-cancer treatments prescribed following this date unless a recurrence occurred. This allowed for a relatively simple evaluation of the other datasets of participants with known recurrence to see if a metastatic ICD-10 code or a new anti-cancer treatment had been documented in the registry datasets to show recurrence. This increased the number of recurrences identified greatly in the registry datasets and could capture nearly all participants' trial recurrences within the registry data that were reported in the trial. It is noted that for prostate cancer percentage recurrences when all the trial databases were used was lower than other tumour types. This may be due to high use of hormone deprivation in recurrence setting which may not be picked up in the hospital records as usually prescribed by the GP. Also PSA failure which is an important marker for prostate cancer recurrence used in trials is deemed not possible to pick up in the EHR records at present.

This method of recurrence capture in EHR would need to be validated in an independent dataset and it was not within the remit of this project to create or validate algorithms for recurrence however it gives strong support to the use of these in the future as other countries have done (163-166). These algorithms could be used as a trigger for the CTU and then verified with the site. If regular extracts were provided by NCRAS then possibly more recurrences could be picked up during the course of the trial which could be used for interim analysis. As mentioned above this method does have its limitations as patients prescribed new hormonal therapies for breast or prostate cancer may not be captured by SACT database. However, this may be possible in the future with GP prescription data looking at new hormonal medication prescribed by the GP.

Adverse outcome data showed low level of concordance between trial data and HES APC data for all the events that were reviewed. This was not reviewed in detail as in the previous chapter of this thesis but again it shows that HES APC data does not collect all adverse events that are important to the trial. The registry, as seen in chapter 1, may pick up some events the trial does not particularly in those lost to follow-up. However, the registry data has likely high false positive rate as the ICD-10 codes cannot match the trial outcomes exactly. The possible false positive rates may be due to broad ICD-10 code definitions which on review are non-serious or incidental findings. In particular with bleeds the ICD10 K922 (gastrointestinal haemorrhage unspecified) is likely to be too non-specific for a bleeding event. This made up most of the bleeding events that were seen in HES and not the trial data. Therefore, HES APC cannot be the only source to find adverse events and may even cause more work to the trial team to follow-up on any possible protocol defined event if there are many false positives. Therefore, for acute events the normal SAE reporting may still be the best way to record these events as per the protocol. However, in chronic disease it may provide a good double check to make sure these conditions are not missed in the final analysis for conditions like dementia, diabetes and macular degeneration.

Staging data was not one of the primary or secondary outcomes of the trial but a variable within registry data that could be useful for many trials either in the recruitment of patients in oncology trials or outcome data for a trial that had stage of primary diagnosis of cancer as the main outcome. This has been compared previously by different trials in breast cancer and prostate cancer with English cancer registry data.

Both showed that either missing data or 'best' stage from multiple sources made the comparability of around half or less (67, 153). This analysis using a cohort of the four different tumour types demonstrates good concordance between the EHR and trial datasets. There is variability between the tumour groups depending on the complexity of the treatment pathway especially with neo-adjuvant treatment in concordance between TNM best and trial data. In colorectal cancer where the pathway can be relatively simple, when radiotherapy is removed, there is a concordance of over 97% whereas gastro-oesophageal has concordance of around 80%. When there has been surgery the histology registry staging usually has good concordance with post-surgical pathology in the trial with concordance of >94%. However, there is more missing data here and not every tumour has surgery as a definitive treatment. Therefore, it is still the case that when there is some subjectivity of assimilating data from various sources for complex cancer treatment pathways this can cause a significant difference between the trial data and registry data.

These findings have notable implications in that when using registry data it is important to understand the concordance of data (before use) based on the required outcomes and also registry used to ensure that it is fit for the purpose required by the trial. Depending on the outcome reviewed the EHR data could possibly be used to replace trial data or be used as a trigger tool for the MRC CTU to check with sites to make sure that trial outcomes are collected with the best accuracy possible.

On review of the feasibility of using this data the cost is relatively low at approximately £400 a year once all participants are recruited and would provide the potential to supplement and improve conventional long-term follow-up. This would seem sufficient to consider its use in the Add-Aspirin as an on-going strategy. Importantly the cost of this though may change in the future depending on which organisation is the data controller which is out of the CTU's control and may be a risk to the trial planned budget if contracts and payments need to be renewed for the data.

However, a potential greater concern is the time lag in the some of the registries particularly the cancer registry with a delay of approximately 18 months and how this would affect the overall analysis if they were to be used. However, in the future NCRAS will release a rapid cancer registration dataset which could help with this. The accuracy of the data may not be as good as the current cancer registry however it could give

information on a likely new cancer or on new treatments such as surgery/radiotherapy or chemotherapy which could act as a trigger to go back to site to validate the detail with sites (167). Also the time lag of data from sites can also be prolonged as seen with the mortality data where death data can take over 6 months to ascertain. This can be due to the frequency of follow-up which is 6 monthly in this trial. This can mean that the site would be unaware of the death for 6 months after the death. There is also the time taken to complete the CRF and also for the data to be entered on to the CTU database. The registry time lag, therefore, maybe not be such a significant issue especially with the lag time in registry data potentially reducing in the future.

There were limitations to this study with the most challenging being the loss of a lot of patients due to incorrect identifiable data, particularly NHS number, that made matching and receiving data from NCRAS difficult. This meant that approximately 10% of participants did not have registry data to compare to. This is under investigation into how this can be improved if Add-Aspirin were to ask for further updated data on this cohort and to make sure new patients have the right details. This is challenging due to the security measures required in transfer and storage of this data to maintain confidentiality – e.g. CTU trial team do not have access to it. There is also an issue for the future with patients not consenting to their EHR data use which consisted of 1.8% of this cohort. This could also be a major problem for CTUs if they were to use this method of collecting data in the future.

This is also a subset of patients from the trial itself even though England has the largest population of participants this study only addresses EHR records that involve them. There is no evidence to say that Scottish, Welsh or Northern Irish data would be any different in the quality to the English data, however, the take up of different datasets is variable with an example of the radiotherapy dataset not covering every hospital in Scotland (168). They also do not necessarily have the same structure or data fields meaning that interpretation may be different and need further analysis. There is no appropriate EHR in India so using registry data only for these English patients may cause a systematic bias in the analysis of the trial which needs to be considered. It needs to be considered whether collecting data from other possible UK registries is of sufficient importance versus the financial and human resource implication needed to apply and process that data. Also if EHR was only used for English participants would this create a systematic bias within this cohort of patients.

A further limitation is the different censorship of datasets between the trial and different cancer registries. Some may argue that this is not a true comparison as the data freeze dates are not the same and therefore you cannot make fair judgements on the comparability of the data. This can be seen within the cancer registry where there is a significant time lag within the data compared to the other datasets. It could also be said that the trial data had a time advantage as well. This analysis shows that all the datasets had inaccuracies and that, apart from the known data cut off points, time did not seem to affect what caused the missing or inaccurate data.

3.5 Conclusion

This was a feasibility study to assess if EHR data could be used for long-term follow-up. This study demonstrates that registry death data could possibly be used alone to ascertain mortality data. However, all outcomes of the trial could not be collected solely by registry data due to inaccuracies and potential lag time of the data. This method of follow-up, at time of writing, represents a relatively low-cost solution to help with long-term follow-up. It could therefore be a valuable resource to assist in collecting primary outcomes for those lost to follow-up at site. It could also supplement conventional long-term follow-up providing some data earlier or to decrease missed events which could be clarified by the site. This would support the long-term follow-up approach in Add-Aspirin to strip back and then decrease the frequency of follow-up CRFs. This study has shown that registry data is continuing to improve and has an important role in trial long-term follow-up but care is needed when using it to make sure that it is validated for the outcome which the trial is investigating. The price and application for this data may also change in the future with a different organisation being the data controller. This may have financial and logistical repercussions for how useful this data is.

4 A prospective cohort study within the United Kingdom Clinical Trial of Ovarian Cancer Screening (UKCTOCS): Aspirin use and cancer incidence

4.1 Introduction

4.1.1 Pragmatic randomised controlled trials

Traditional RCTs have been described by critics as expensive, unrepresentative of the general population, and subject to loss of follow-up and potential biases (13). Despite this RCTs are still the gold standard for investigating a new medical intervention. EHRs have been suggested as a resource to be used in 'pragmatic' RCTs to try to address some of these issues. This has been achieved in other countries with the 'TASTE' trial as an example which worked to great success (52).

A British example is the UK SALFORD Lung study (NCT01551758) designed to evaluate the effectiveness of once daily 100 micrograms or 200 micrograms fluticasone furoate with 25 micrograms of vilanterol or optimised usual care in symptomatic asthma or chronic obstructive pulmonary disorder (COPD) (27). This study used an established electronic medical record (EMR) system which connected primary and secondary care to follow-up the patient with minimal change to the patients' normal follow-up. A "change of culture" was needed to help this trial recruit participants within a GP practice due to their lack of experience of staff in trials but the pragmatic follow-up method was effective. Results from the study demonstrated that the intervention improved asthma control and did not increase SAEs (169, 170).

Within the MRC CTU at UCL there has been interest in designing a pragmatic trial using EHR to facilitate trial conduct to assess aspirin in the primary prevention setting in those at a high risk of developing cancer. Primary prevention trials often require a large cohort with a long follow-up period. A pragmatic trial using EHR has the potential to decrease trial costs, facilitate recruitment of a large diverse population and improve long-term follow-up data collection.

Therefore, to further evaluate this potential project a cohort study using United Kingdom Trial of Ovarian Cancer Screening trial (UKCTOCS; ISRCTN 22488978) was proposed as the trial had access to national English and Welsh registry EHR data and also 'aspirin use' information from a questionnaire that participants completed. The objective of the study was to gain experience of working with the relevant EHR data for the proposed future trial and also to add to the evidence base for aspirin as a pharmacological primary prevention agent for cancer. The evidence for aspirin as a primary prevention agent against cancer is growing both in terms of basic science and observational studies but RCT evidence is more controversial.

4.1.2 Evidence for aspirin in primary prevention

Laboratory evidence supports aspirin as a chemopreventive agent but the exact mechanism of action is still under debate. Aspirin is known to directly inhibit the enzyme cyclo-oxygenase (COX) (otherwise known as prostaglandin-endoperoxide synthase (PTGS0)), of which there are two isoforms, COX-1 and COX-2. COX catalyses the conversion of arachidonic acid to prostaglandins and other downstream inflammatory mediators including thromboxane, which play a role in immune modulation, cell proliferation, control of apoptosis and tumour growth (171, 172).

The mechanism of action of aspirin in cancer is predominately thought to be an anti-platelet effect via inhibition of COX-1 acetylation in platelets (173, 174). *In vivo* studies have demonstrated that platelet inhibition decreased the metastatic potential of carcinoma cells by decreased thromboxane A2 and prostaglandin (PGE2) (175).

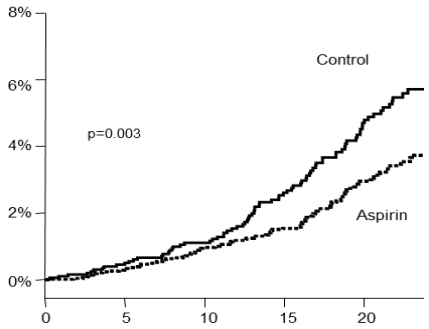
Aspirin's possible direct inhibition of COX-2 and subsequently prostaglandin production acts to block immunosuppression, tumour immune evasion and retards tumour growth *in vivo* (176). The evidence of aspirin in primary prevention is strongest in colorectal cancer. Here COX-2 has been shown to be significantly over-expressed in the majority of colorectal carcinomas and the Nurses' Health Study and the Health Professionals Follow-Up Study (HPFS) demonstrated that regular aspirin conferred a significant risk reduction in colorectal cancers over-expressing COX-2 (RR 0.64, 95% CI 0.52-0.78) (177).

Bosetti *et al*. in their meta-analysis summarised data from case-control and cohort studies, involving 737,409 cases, showed that aspirin use was associated with a significant decrease in the risk of overall cancer (RR 0.89, 95% CI 0.87-0.91) (178). Recently they have updated this review demonstrating ongoing benefit for all digestive tract cancers apart from head and neck but particularly colorectal cancer (RR 0.73, 95% CI 0.69–0.78) (179). The benefit of aspirin in the primary prevention setting has also been supported in similar meta-analysis of observational studies (180, 181).

The evidence for aspirin in the primary prevention setting is more controversial when RCTs are reviewed. Data from a meta-analysis of RCTs by Rothwell *et al*., from aspirin primary prevention studies in cardiovascular disease, published in the early 2010s, demonstrated improvements in cancer incidence from 3 years onwards (OR 0.76, 95% CI 0.66-0.88) and also a reduction in cancer mortality after 5 years (OR 0.63, 95% CI 0.49-0.82) (182). An updated meta-analysis by Wu *et al.* published in 2020 incorporated more recent large studies that have predominately been negative for aspirin benefit in cancer prevention (183-185). This meta-analysis demonstrated no association with aspirin and decreased cancer incidence (RR 1.01, 95% CI 0.97 to 1.04, p= 0.72), and total cancer mortality (RR 1.00, 95% CI 0.93 to 1.07, p= 0.90) (186). The recent negative studies that have been included have a relatively short follow-up time. It is widely considered that the primary preventative effects of aspirin may take 10-20 years to become apparent as demonstrated by these graphs of different trials follow-up in of aspirin’s effect on risk of colorectal cancer after long-term follow-up.

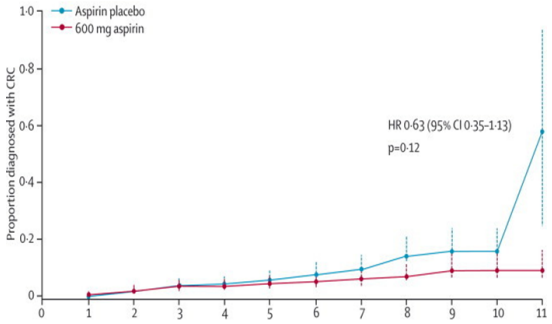
Figure 4-1: Effect of aspirin on risk of colorectal cancer after long-term follow summaries graphs from different trials (187-190)

BDAT & UK-TIA



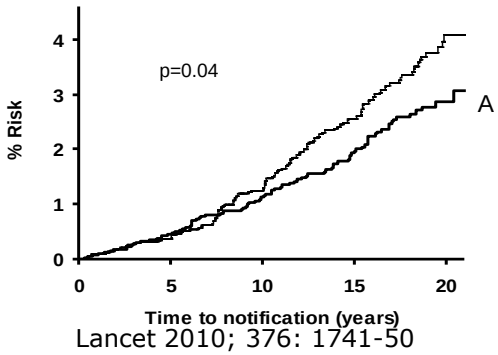
Lancet 2007; 369:1603-13

CAPP2



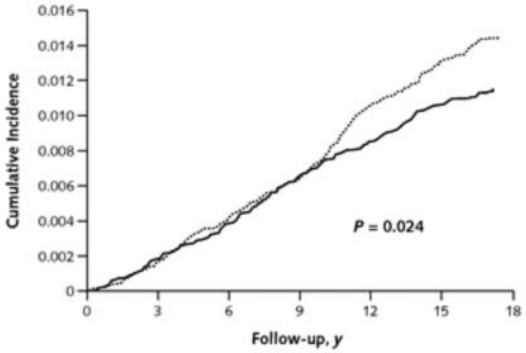
Lancet 2011; 378:2081-2087

TPT & SALT



Lancet 2010; 376: 1741-50

WHS



Ann Intern Med 2013

Rothwell *et al.* have demonstrated that, with 20 year median long-term follow-up, an increased scheduled duration of aspirin of more than 5 years had a further incremental beneficial effect on the risk of cancer death and this was sustained with the extended follow-up (189). Therefore, the negative result of Wu *et al*. meta-analysis could reflect the relatively short median study follow-up of less than 5 years. This potential criticism was addressed in their paper as even when they looked at length of aspirin use up to and greater than 10 years they saw no benefit of aspirin on cancer incidence and survival. The most encouraging data in observational studies and Rothwell's analysis were mainly in colorectal cancer/digestive tract cancers and as Wu *et al.* did not differentiate between tumour type this maybe another reason why no benefit was seen (191, 192). From randomised trial data the benefit of aspirin is still debatable and this has been reflected in US Preventative services task force (USPSTF) most recent published draft guidelines. They have modified their opinion and no longer advise aspirin as a primary preventative medication for cardiovascular disease or colorectal cancer over the age of 60 (193). Therefore, a primary prevention trial evaluating aspirin assessing cancer incidence and survival would be an important question to try to answer.

4.1.3 UKCTOCS trial

UKCTOCS was a RCT aiming to assess the impact of screening on mortality from ovarian cancer. They recruited 202,638 postmenopausal women, aged 50-74, from 2001-2005 (60) and found that there was no reduction in ovarian cancer deaths in the screening group (194). The study used EHRs to follow-up the participants, and women consented for these records to be used in secondary studies. UKCTOCS continued to receive data on participants through the cancer registry, hospital admission data and ONS mortality data in England and Wales to the end of the trial in 2020. During the follow-up period two questionnaires were sent to participants. In the second questionnaire administered in 2014 data about aspirin use and timing of use was collected. Therefore, a study was devised and ethically approved to investigate the hypothesis that aspirin decreases the risk of developing cancer using data from the UKCTOCS cohort.

4.1.4 Prospective cohort studies in aspirin primary prevention

Similar studies have used similar research methodology and employed EHR registry data. This approach has been extensively used in Scandinavia where they have excellent national prescription data (195). However other long-term cohort studies have used participant questionnaires to assess aspirin's role in primary prevention. The best examples are the Health Professionals Follow-up Study (HPFS), Nurses' Health Study and the Prostate, Lung, Colorectal and Ovarian (PLCO) Screening Study (NCT00002540). HPFS and Nurses' Health Study were two large cohort studies based in the USA. HPFS recruited 51,529 male health professionals, age 40-75, who returned a mail health questionnaire in 1986 (196). The Nurses' Health Study recruited 121,700 female registered nurses, aged 30 to 55, in 1976 (197). Questionnaires included information on aspirin use, diet and cancer diagnosis. In both studies questionnaires were sent out every 2 years up with detailed questions on aspirin use and dose in the preceding two years.

The PLCO screening study was a RCT to determine if screening investigations in prostate, lung, colorectal and ovarian cancer can reduce mortality in that cancer. It included 154,901 participants, aged 55-74, from US screening centres that recruited participants between 1993 and 2001. The PLCO study had less extensive data on aspirin use with a baseline questionnaire asking about aspirin use and then a follow-up questionnaire between 2006-2008 (198). 101,098 participants completed the supplementary questionnaire in 2006-2008. The analysis was based on all those who completed the baseline questionnaire so long-term aspirin data was not known for approximately a third of the participants.

The 3 studies above have investigated whether the use of aspirin decreases the risk of developing cancer, they confirm the evidence from other observational studies, that aspirin use is associated with a reduction in cancer risk particularly for some cancer types. (196-200) There are some strengths and limitations of these studies. They all had a very good system for collecting cancer occurrence and are large cohorts with long-term follow. HPFS and Nurses' Health Study, however, have more information on aspirin use and the dosing than the PLCO screening study.

4.2 Methodology

4.2.1 Primary research question

Is aspirin use associated with a reduced risk of developing a primary cancer within the cohort of participants in the UKCTOCS trial?

4.2.2 Secondary research question

1. Within this cohort of participants who develop cancer is aspirin use associated with improved cancer outcomes measured as cancer mortality?
2. Within this cohort of participants is aspirin use associated with an increased rate of major upper GI and intracranial (CNS) bleeds requiring admission to hospital?
3. Is data from EHRs sufficient to be used for follow-up in a RCTs of aspirin as primary prevention?

4.2.3 Inclusion Criteria

Participants within UKCTOCS must have:

- Consented to secondary research
- Not been recruited in Northern Ireland (due to EHR availability)
- Answered (yes or no) to have the “ever taken low dose aspirin question” on the 2014 follow-up questionnaire
- No history of cancer recorded on the recruitment questionnaire
- No cancer registration prior to randomisation for ICD10 “C*” or ICD09 “140*”-“239*” (cancer diagnosis codes)

4.2.4 Study design

This project was a prospective cohort study using a subset of data which were previously collected from women participating in the UKCTOCS. Within UKCTOCS, a questionnaire was sent to all participants in 2014 asking if they have taken or were currently taking regular aspirin. 83,528 participants filled in the questionnaire with approximately 50,000 participants answering the question regarding aspirin use. This questionnaire data was combined with participants' national registry data that holds records on cancer diagnosis, treatment and death.

Data from English and Welsh cancer registries, NHS Digital English HES data/ Welsh equivalent PEDW data and ONS mortality data was interrogated for each participant in this cohort. Data regarding any cancer diagnosis and date was extracted. Cancer outcome data was dependent on registry information availability. Cancer outcomes were based on cancer mortality from ONS mortality data.

The rate of upper GI and CNS bleeds requiring admission to hospital was assessed using NHS Digital HES APC data and Welsh PEDW data. These were assessed using pre-specified ICD10 diagnosis codes.

4.2.5 Aspirin ascertainment

In the 2014 questionnaire participants answered "Yes" or "No" to the question 'Have you ever taken the following medication: Low dose Aspirin'. Data was also collected on start and stop dates. This question was also asked for tamoxifen and statins.

On review of the data by the trial data manager Andy Ryan (AR) there were a significant number of questionnaires where the participant had not stated the date when they completed the questionnaire. In this case an average date of questionnaire completion was used which was 12/04/2014. This was important for potentially considering the start and stop dates for aspirin use. 'Nonsense' questionnaire completion dates that were on the database were corrected using the original scanned questionnaires by myself (AM) and AR.

Further to this aspirin start and stop dates also had missing information or “nonsense” dates. Out of the 12,432 participants who answered “Yes” to the aspirin use question, there were start dates for 7,881 of the participants and 2,117 stop dates. Stop dates are expected to be fewer as aspirin started for cardiovascular protection is usually/often a lifelong intervention. “Nonsense” start and stop dates on the database were corrected using the original scanned questionnaires by AM. Rules were established to create an appropriate approximate date for dates where partial data was missing (day or month). If day was missing then the 15th day of the month was used and if month was missing then June was used e.g., if both missing but year stated as 1986, it would be 15/06/1986.

The initial analysis was planned based on whether the participant stated “Yes” to the aspirin use questionnaire versus if a participant stated “No”. The effect of duration of aspirin use was considered an important aspect of this analysis. As there was significant missing start and stop dates, an imputation model was used to model start and stop dates for those without data which is discussed in more detail in the statistical analysis section below.

4.2.6 Cancer ascertainment

Cancer censorship dates were not the same throughout the cohort due to data available from the cancer registry in England and Wales. In England the censorship date was 1st September 2018 and Wales 1st August 2016. The majority of patients in this cohort were based in England.

Cancer data included all ICD-10 codes with “C*”. For the three common cancers the codes used were Breast- C50*, Lung C34* and Colorectal C18*, C19* and C20*. Codes that inferred precancerous or benign disease starting with “D*” were excluded from the analysis.

4.2.7 Cancer Mortality

During the initial concept of the study several outcome ‘proxy’s’ for aspirin benefit on cancer outcome were considered. These outcome ‘proxy’s’ were either looking at metastatic disease development, or treatments which inferred localised disease rather than metastatic or cancer death. It was apparent after review of the data that either data was not sufficient or too complex for the scope of this project to use metastatic disease or treatment as a ‘proxy’ for cancer outcome. Therefore, in a protocol amendment it was stated that cancer mortality alone would be the secondary outcome measure.

Mortality data was gathered from ONS mortality data. This dataset was censored on 18/09/2020. A cancer death was defined as any death coded with ICD-10 code “C*”. All deaths which had an ICD-10 “C*” code as any of the causes of death was included in the analysis. This was due to the fact that often despite advanced cancer other causes like ‘pneumonia’ or ‘cardiovascular disease’ are stated as the primary cause of death on the death certificate.

4.2.8 Major upper GI and CNS haemorrhage ascertainment

NHS Digital HES APC data and Welsh PEDW data was used to assess major upper GI haemorrhage and intracerebral major haemorrhage requiring hospital admission. This data was censored on the 05/06/2020 in the English dataset and on 14/07/2020 for the Welsh data. An event was defined as a participant being admitted into hospital for the codes in the tables below which were established codes from previous publications and the particular codes are within the **Appendix D** (161).

Due to the nature of admission data from England and Wales only the first event was used from diagnosis box 1 (primary cause of admission) to create an event for statistical analysis. This decreased the false positive reasons for admission and also excluded repetition of the codes despite admission for another reason which can be a significant problem as discussed in chapter 1 pg. 78.

4.2.9 Ethics

UKCTOCS was approved by the UK North West Multicentre Research Ethics Committees (North West MREC 00/8/34) with site specific approval from the local regional ethics committees and the Caldicott guardians (data controllers) of the primary care trusts. Women who did not provide consent to allow their data to be used for secondary studies were excluded from this study.

This study used data that had already been collected from volunteers participating in the UKCTOCS trial. Participants provided written informed consent which allowed the research team to access their medical notes for information relevant to the research and their data to be used in ethically approved secondary studies:

“I give permission for individuals from the UKCTOCS research team to access my medical notes for information relevant to the research. I understand that regulatory authorities may also access this information.”

Since 2015, Section-251 approval from the Confidentiality Advisory Group had also been obtained that allows processing of confidential patient information without consent.

This study ‘Aspirin use and cancer incidence in UKCTOCS’ was approved by London Hampstead research ethics committee on 18/12/2019 (Ref No 19/LO/1989) and by UCL REC on 13/02/2020. An amendment was approved by the UCL REC committee on 13/08/2020 with minor amendments to the protocol and approval for the project to be used for educational purposes.

4.2.10 Data Governance

The identification of the participants who were eligible for this sub-study was carried out by core researchers on the UKCTOCS team in the Gynaecological Cancer Research Centre. All identifiable information was pseudo-anonymised and checked by the UKCTOCS Data Manager before data was released.

Identifiable data previously collected from and about UKCTOCS participants was stored in the UCL data safe haven. The data safe haven has been certified to the ISO27001 information security standard and conforms to NHS Digital's Data Security and Protection Toolkit. It was built using a walled garden approach, where the data is stored, processed and managed within the security of the system, avoiding the complexity of assured end point encryption. A file transfer mechanism enables information to be transferred into the walled garden simply and securely. Only core UKCTOCS researchers were authorised to access the data using individualised login/passwords.

No patient identifiable information was provided to anyone outside the core UKCTOCS team at any stage of this project. All analysis took place under a secure environment within the UCL data safe haven and all information was pseudo-anonymised with conversion to unique study number and age used for analysis rather than date of birth. When summary data/ analysis was shared with anyone outside the core team all identifiable information was removed and the files were checked by the UKCTOCS Data Manager prior to any release.

4.2.11 Statistical Analysis

Sample size was based on the number of participants who answered the question as to whether they took aspirin in the 2014 questionnaire. Analysis of data was based on the 47,449 participants who answered the question. Analysis time was from age at randomisation to age at cancer diagnosis, with censorship to the most recent UKCTOCS data available at the time of analysis, unless the individual left the NHS or died of a non-cancer cause before then. The statistical analysis plan was devised by AM and Matthew Burnell (MB), UKCTOCS statistician and all statistical analysis was done by MB within the UCL data safe haven.

For the primary analysis a basic cox regression model between aspirin ‘Yes’ versus ‘No’ was initially used. The cox regression models for survival assessed specifically the nature of any relationship between aspirin use and (reduced) cancer risk. The model was adjusted for potential confounders which were asked at baseline and in the

questionnaire in 2014. This is included Body Mass Index (BMI), smoking, alcohol, hypertension, cardiac/stroke history, high cholesterol, diabetes, medication (statin use and hormone replacement therapy), age at menarche, hysterectomy and family history of cancer.

The aspirin exposure variable attempted to reflect the duration of aspirin use as well the timing of use relative to (possible) cancer diagnosis. Hence the variable was time-varying and incorporated cumulative use. That was, how risk might change with length of aspirin use and also with length of time since last aspirin use. The model was adjusted for the same potential confounders.

After initial preliminary analysis there was significant missing data particularly in the aspirin date of use variable. Different statistical strategies were therefore used to account for this missing data to achieve the most accurate and correct result for the data that was available. For ease of narrative two cohorts of participants were defined cohort ‘A’ and ‘B’. In cohort A participants answered yes to the aspirin use question in 2014 and could have had a cancer diagnosed anytime between randomisation and the date of analysis whilst in cohort B they answered yes to the aspirin question and their cancer was diagnosed after 2014. The advantage of cohort B is that there is certainty that the individual was taking aspirin prior to their cancer diagnosis. The main disadvantage is that the follow-up time is relatively short (2014 to 2018 for English participants and 2014-2016 for Welsh participants) which is important given that the effects of aspirin on cancer incidence are long-term i.e. after a period of 5-10 years. It also reduced the number of incident cancers in cohort B compared to cohort A thereby reducing the power of the study. It is also of note that aspirin use once started is usually life-long therefore the data from cohort A was considered to be of value and presented alongside data from cohort B.

In cohort ‘A’ data from randomisation was used as above with a multiple imputation strategy. Specifically chained equations were used in order to multiply impute data to create 20 complete datasets under the assumption of ‘Missing At Random’. Values were imputed for data missing in the potential confounders but also specifically for the missing start and stop dates for aspirin use. Here, the chained equation was a truncated regression model, conditional on aspirin use, with lower and upper bounds defined by age at randomisation and age at censorship in the absence of actual start

and stop dates. The analysis results from the 20 datasets were combined using Rubin's rules to create a survival model with correct standard errors.

The approach described above, rather than the basic cox model, was also used for the secondary analysis assessing the effect of aspirin on the incidence of certain defined common cancers (breast, colorectal and lung) and the effect on all cancer mortality and individual cancer mortality.

In cohort 'B' an additional strategy was also used in addition to the above multiple imputation method to negate the particular issue of missing aspirin start dates. The participants 'aspirin use' was gathered from a questionnaire in 2014. Aspirin use at the 2014 questionnaire was definite even if they have not filled in aspirin start date. This would make sure that participants had started their aspirin prior to a cancer diagnosis. Therefore, analysis was performed as above using initially a basic cox regression for aspirin "Yes" versus "No" to assess effect on all cancer incidence and all cancer mortality but analysis time was from age at completion of questionnaire to age at cancer diagnosis, with censorship to the most recent UKCTOCS data available at the time of analysis, unless the individual left the NHS or died of non-cancer cause before then.

Lastly the same survival analysis was carried out to account for aspirin length of use with the same imputed data but with the analysis start at age of participant at completion of questionnaire. This strategy was also used for the incidence of certain defined common cancers (breast, colorectal and lung) and that certain cancer mortality.

The other secondary analysis of risk of admission to hospital with a major upper GI or CNS haemorrhage was only analysed in cohort B using basic cox regression model. This was based on the fact that event numbers were low and therefore the most suitable statistical method was thought to be aspirin 'Yes' versus 'No' with no cumulative analysis starting from age at completion of 2014 questionnaire. The same confounders were used for this analysis due to the significant overlap of risk on defined event.

4.3 Results

4.3.1 Primary cohort analysis from recruitment cohort ‘A’

47,449 participants were available for analysis from the original 202,638 UKCTOCS cohort following pre-defined exclusion criteria demonstrated in the consort diagram (**Figure 4.2**). Since recruitment began (i.e. cohort A) there were 6,992 cancers that developed during follow-up until censorship of cancer registration data. The baseline characteristics of the cohort are shown in **Table 4.1**. Aspirin users were slightly older mean 62 years versus 59 years at trial entry and more likely to have a history of cardiovascular disease, hypertension, stroke, hypercholesterolaemia, diabetes and be taking a statin. This was to be expected as aspirin is usually prescribed for primary or secondary prevention of cardiovascular disease and often co-prescribed with a statin. Other potential confounders had a similar distribution between the two groups including family history of breast or ovarian cancer and hormone replacement use.

Figure 4-2: Consort diagram of participants available for analysis from original UKCTOCS cohort

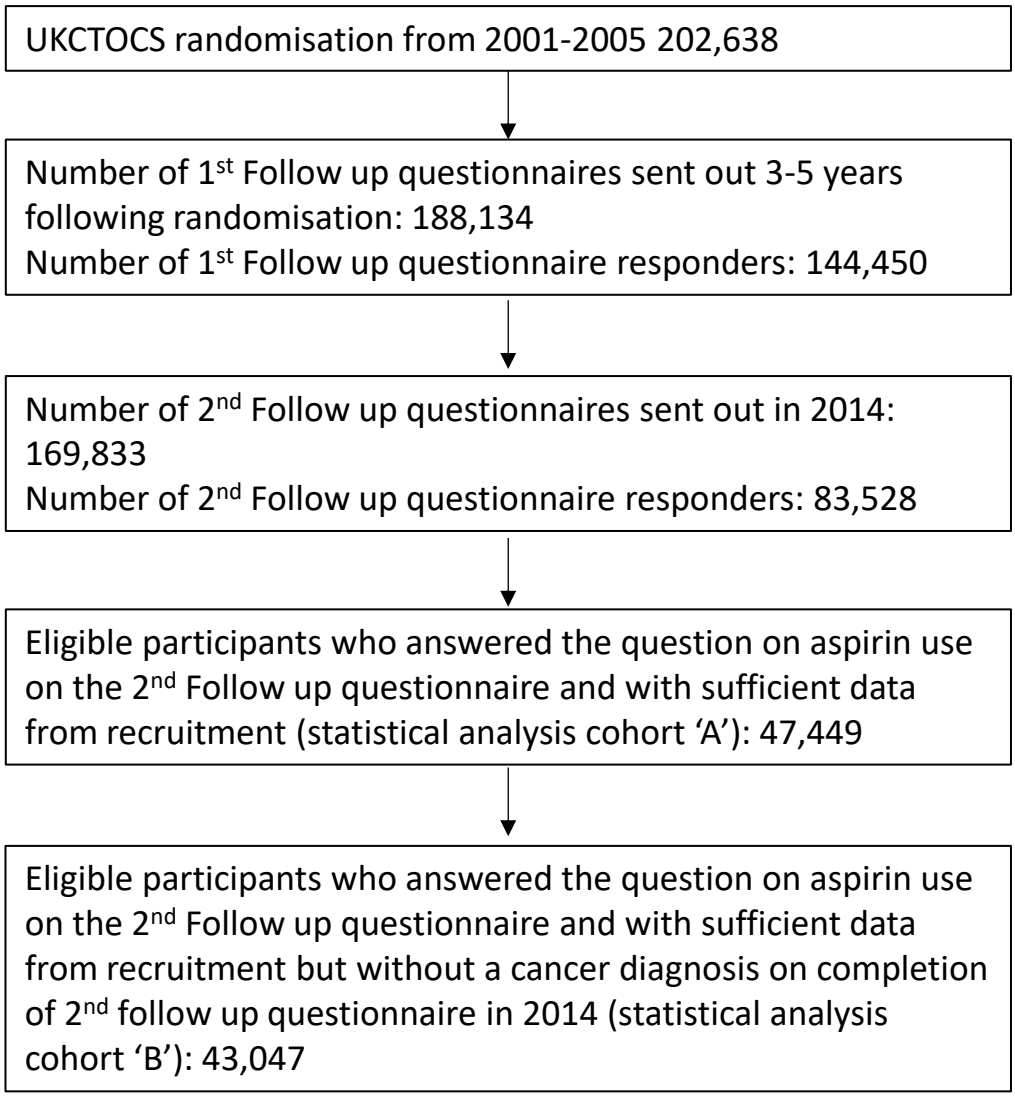


Table 4-1: Participant characteristics at baseline and updated after questionnaire 2

Parameter	Aspirin No	Aspirin Yes	Chi squared test (categorical) or two sample t-test (continuous variables)
Participants baseline questionnaire information	34,806	12,643	N/A
Age at recruitment/ Mean	59 (Range 50-74)	62 (Range 50-74)	p=<0.001
BMI/ Mean	26	27	p=<0.001
Alcohol/ Median number of units a week	1-3	1-3	p=<0.001
Never Smoker (%)	19,088 (70)	6,489 (66)	p=<0.001
Hypertension (%)	6,443 (18)	5,682 (45)	p=<0.001
Cardiac disease (%)	392 (1)	1,854 (15)	p=<0.001
Previous Stroke (%)	117 (0.3)	493 (4)	p=<0.001
High Cholesterol (%)	2,933 (8)	4,953 (39)	p=<0.001
Diabetes/ %	481 (1)	1,213 (10)	p=<0.001
Hormone replacement therapy (%)	6,999 (20)	2,374 (19)	p=<0.001

Hysterectomy (%)	5,016 (14)	2,407 (19)	p=<0.001
Family history of: Ovarian cancer (%) Breast Cancer (%)	 540 (2) 2,350 (7)	 192 (2) 814 (6)	 p=0.798 p=0.226
Mean age at Menarche	12.9	12.9	p=0.140
Mean age at last period	49.5	49.2	p=<0.001
Questionnaire 2 updated information			
Age at questionnaire 2/ Mean	70 (Range 59-87)	73 (Range 59-87)	p=<0.001
Hypertension (%)	9,262 (27)	7,330 (58)	p=<0.001
Cardiac Disease/ %	822 (2)	3,337 (26)	p=<0.001
Previous Stroke/ %	425 (1)	974 (8)	p=<0.001
Diabetes/ %	1,228 (4)	2,029 (16)	p=<0.001
High Cholesterol/ %	3,376 (10)	5,740 (45)	p=<0.001
Statin Use/ %	2,768 (8)	8,581 (86)	p=<0.001

The primary analysis was to assess if cancer incidence was decreased with aspirin use (**Table 4.2**). When aspirin use “Yes” versus “No” was assessed in all participants from recruitment in a basic cox model there was no significant difference (HR 1.01 95% CI 0.96-1.06 p=0.680) and when this model was adjusted for the above baseline and questionnaire 2 confounders in **Table 4.1** there was still no significant difference between the two (HR 0.93 95% CI 0.85-1.01 p=0.087). Due to poor data on start and stop dates for aspirin use an imputation model was used to assess whether there was an effect of duration of aspirin use on cancer incidence in a further survival analysis. This analysis of cohort ‘A’, which took into consideration cumulative use of aspirin, also showed no difference in cancer incidence versus no aspirin use (HR 1.00 95% CI 1.00-1.01 p=0.137) in an unadjusted model and when adjusted for confounders this made no difference to the outcome (HR 1.00 95% CI 0.99-1.01 p=0.723).

However, when this analysis was also repeated for separate common cancers breast, lung and colorectal. The same analysis with the imputation dataset was used for each cancer to assess if the cumulative use of aspirin decreased the risk of that individual cancer. In total there were a total of 1,700 breast, 481 colorectal and 256 lung cancers. When analysis on cancer incidence was done on each cancer there was a significant decrease in cancer incidence after a cumulative aspirin analysis using the same imputation dataset for breast (HR 0.91 95% CI 0.88-0.94 p=<0.0001), colorectal (HR 0.91 95% CI 0.86-0.96 p=0.001) but not in lung cancer cohort (HR 0.99 95% CI 0.95-1.04 p=0.739) before adjusting for confounders. When the model was adjusted for confounders then a significant result was also found with a decrease in cancer incidence in breast (HR 0.86 95% CI 0.82-0.89 p=<0.0001) and colorectal (HR 0.87 95% CI 0.81-0.94 p=<0.0001) cancers and in the lung cancer cohort (HR 0.93 95% CI 0.88-0.99 p=0.026). These results are summarised in **Table 4.2**.

Table 4-2: Primary Analysis of cancer incidence in cohort 'A' for all cancers and defined individual cancers

Cancer incidence comparison	Cancer incidence Aspirin "No"	Cancer Incidence Aspirin "Yes"	HR Unadjusted	HR Adjusted
Cohort total	34806	12643	N/A	N/A
All cancer incidence analysis from recruitment				
All cancer incidence using basic cox model from recruitment	4850	2142	HR 1.01 95% CI 0.96-1.06 p=0.680	HR 0.93 95% CI 0.85-1.01 p=0.087
All cancer incidence using imputation data for start and stop dates from recruitment			HR 1.00 95% CI 1.00-1.01 p=0.137	HR 1.00 95% CI 0.99-1.01 p=0.723
Individual cancer incidence using imputation data for start and stop dates from recruitment				
Breast	1217	483	HR 0.91 95% CI 0.88-0.94 p=<0.0001	HR 0.86 95% CI 0.82-0.89 p=<0.0001
Colorectal	338	143	HR 0.91 95% CI 0.86-0.96 p=0.001	HR 0.87 95% CI 0.81-0.94 p=<0.0001
Lung	160	96	HR 0.99 95% CI 0.95-1.04 p=0.739	HR 0.93 95% CI 0.88-0.99 p=0.026

4.3.2 Secondary analysis of cancer mortality from randomisation cohort ‘A’

A secondary analysis of cancer mortality was also carried out based on death certification of any cancer ICD-10 code (**Table 4.3)**. There was a total of 1,406 deaths in total from cancer since recruitment in the cohort A. When aspirin use “Yes” versus aspirin use “No” was assessed using analysis with the same imputed data from randomisation there was no significant difference in cancer mortality from any cancer death (HR 1.00 95% CI 0.99-1.02 p=0.718). This did not change when it was adjusted for confounders (HR 1.02 95% CI 0.99-1.04 p=0.200).

This was also repeated for individual cancer subtypes mortality. There was 127 breast, 131 colorectal and 256 Lung cancer deaths. When the survival analysis was used with the same imputed data from randomisation there was no difference in cancer mortality between the aspirin “Yes” and aspirin “No” cohorts in breast cancer (HR 0.87 95% CI 0.76-1.00 p=0.054), and lung cancer (HR 0.93 95% CI 0.87-1.00 p=0.055) but significance in colorectal cancer (HR 0.81 CI 0.95% 0.68-0.97 p=0.023) unadjusted. When it is adjusted for confounders then the results were significant for all three cohorts with breast cancer (HR 0.81 95% CI 0.67-0.97 p=0.027), colorectal (HR 0.79 95% CI 0.64-0.98 p=0.031) and Lung cancer (HR 0.87 95% CI 0.79-0.97 p=0.009). These results are summarised in **Table 4.3**.

Table 4-3: Secondary analysis of cancer mortality from randomisation cohort 'A'

Cancer mortality comparison	Cancer mortality Aspirin use "No"	Cancer mortality Aspirin use "Yes"	HR Unadjusted	HR Adjusted
Cohort total	34806	12643	N/A	N/A
All cancer mortality analysis from recruitment				
All Cancer mortality using imputed data for aspirin start and stop dates	930	476	HR 1.00 95% CI 0.99-1.02 p=0.718	HR 1.02 95% CI 0.99-1.04 p=0.200
Individual cancer mortality using imputed data for aspirin start and stop dates from recruitment				
Breast	84	43	HR 0.87 95% CI 0.76-1.00 p=0.054	HR 0.81 95% CI 0.67-0.97 p= 0.027
Colorectal	95	36	HR 0.81 95% CI 0.68-0.97 p=0.023	HR 0.79 95% CI 0.64-0.98 p=0.031
Lung	165	91	HR 0.93 95% CI 0.87-1.00 p=0.055	HR 0.87 95% CI 0.79-0.97 p=0.009

4.3.3 Primary analysis of cancer incidence post questionnaire 2 cohort 'B'

Following the review of the data where there was significant missing data particularly of start and stop dates of aspirin alternative statistical methods were used to determine the relationship of aspirin on cancer incidence in this cohort of women. There were 43,047 women available for analysis in cohort B (**Table 4.1**). There was 2,656 cancers diagnosis following questionnaire 2 in 2014.

Firstly, the primary analysis looked at aspirin use "Yes" versus aspirin use "No" on cancer incidence without taking into consideration time on aspirin **(Table 4.4).** In an unadjusted analysis for confounders using a basic cox model there was no significant difference between aspirin use "Yes" and aspirin use "No" (HR 1.01 95% CI 0.96-1.06 p=0.680). This did not change when the analysis was adjusted for confounders (HR 0.93 95% CI 0.81-1.06 p=0.252).

This survival analysis was also repeated with the same imputed data but also to take into account aspirin duration of use **(Table 4.4).** This was to combine both statistical techniques to make sure aspirin was definitely started prior to cancer diagnosis and also to make sure length of aspirin use was reflected in the analysis. In this analysis cancer incidence for aspirin use "Yes" versus aspirin use "No" there was a significant difference between the two cohorts (HR 1.02 95% CI 1.01-1.023 p=0.003) in favour of aspirin use "No" which was not maintained when adjusted for confounders (HR 1.01 95% CI 0.99-1.03 p=0.064).

This survival analysis was again repeated for breast, colorectal and lung cancer. This demonstrated no significant difference in breast or colorectal cancer when unadjusted (HR 1.00 95% CI 0.98-1.03 p=0.854) (HR 1.01 95% CI 0.97-1.04 p=0.700) and adjusted for confounders (HR 1.01 95% CI 0.98-1.04 p=0.394) (HR 1.00 95% CI 0.96-1.05 p=0.964). In lung cancer there was a benefit of aspirin use "No" (HR 1.06 95% CI 1.03-1.09 p=<0.0001) in the unadjusted analysis which was just maintained in the adjusted analysis (HR 1.04 95% CI 1.00-1.08 p=0.034).

Table 4-4: Primary Analysis of cancer incidence in cohort 'B' for all cancers and defined individual cancers

Cancer incidence comparison	Cancer incidence Aspirin "No"	Cancer Incidence Aspirin "Yes"	HR Unadjusted	HR Adjusted
Cohort total	31772	11275	N/A	N/A
Cancer incidence analysis basic cox model	1860	796	HR 1.01 95% CI 0.96-1.06 p=0.680	HR 0.93 95% CI 0.81-1.06 p=0.252
All cancer incidence using imputation data for start and stop dates			HR 1.02 95% CI 1.01-1.02 p=0.003	HR 1.01 95% CI 0.99-1.03 p=0.064
Individual cancer incidence only from Questionnaire 2 using imputation data for start and stop dates				
Breast	434	136	HR 1.00 95% CI 0.98-1.03 p=0.854	HR 1.01 95% CI 0.98-1.04 p=0.394
Colorectal	125	58	HR 1.01 95% CI 0.97-1.04 p=0.700	HR 1.00 95% CI 0.96-1.05 p=0.964
Lung	100	72	HR 1.06 95% CI 1.03- 1.09 p=<0.0001	HR 1.04 95% CI 1.00-1.08 p=0.034

4.3.4 Secondary analysis of cancer mortality and upper GI/CNS major haemorrhage post Questionnaire 2

The same two analysis were repeated for the secondary analysis of cancer mortality in any cancers and specific cancers which occurred following questionnaire 2 in 2014. There were 1,406 cancer deaths following the 2014 questionnaire 2 **(Table 4.5)**. There was no statistically significant difference in cancer death unadjusted or adjusted for confounders (unadjusted HR 1.03 95% CI 0.92-1.16 p=0.560, adjusted HR 0.99 95% CI 0.82-1.18 p=0.874) and this did not change when survival analysis took into account aspirin duration of use on cancer mortality (unadjusted HR 1.00 95% CI 0.96-1.02 p=0.670, adjusted HR 1.02 95% CI 0.99-1.04 p=0.190). Cancer deaths in individual cancers also assessed with no significant difference between aspirin use in any of the cohorts of breast, colorectal and lung cancer.

Finally Upper GI/CNS haemorrhage was just assessed against events post questionnaire 2 in 2014 after it was known that the patient definitely had taken the aspirin prior to the event (**Table 4.5)**. This was assessed for aspirin use “Yes” versus aspirin use “No”. There were few events for patients with their primary cause of admission being major upper GI/ CNS haemorrhage using HES data and ICD-10 codes described above at 227 events. There was a statistically higher risk of these major haemorrhage events in those with aspirin use “Yes” versus aspirin use “No” (HR 1.64 95% CI 1.25-2.15 p=<0.0001). When this was adjusted for confounders then the statistical significance was lost but there was still a trend towards increased risk with aspirin use (HR 1.21 95% CI 0.78-1.89 p=0.40).

Table 4-5: Secondary analysis of cancer mortality and upper GI/CNS haemorrhage cohort 'B'

Cancer mortality comparison	Cancer mortality Aspirin use "No"	Cancer mortality Aspirin use "Yes"	HR Unadjusted	HR Adjusted
Cohort total	31772	11275	N/A	N/A
Cancer mortality analysis basic cox model aspirin "Yes" vs "No"	930	476	HR 1.03 95% CI 0.92-1.16 p=0.560	HR 0.99 95% CI 0.82-1.18 p=0.874
Cancer mortality analysis using imputed aspirin start and stop dates			HR 1.00 95% CI 0.96-1.02 p=0.670	HR 1.02 95% CI 0.99-1.04 p=0.190
Individual cancer mortality following Questionnaire 2 using imputed aspirin start and 'stop' dates				
Breast	84	43	HR 1.00 95% CI 0.92-1.07 p=0.910	HR 1.01 95% CI 0.93-1.11 p=0.770
Colorectal	95	36	HR 1.00 95% CI 0.93-1.05 p=0.720	HR 1.00 95% CI 0.93-1.08 p=0.990

Lung	165	91	HR 1.03 95% CI 1.00-1.06 p=0.058	HR 1.02 95% CI 0.98-1.06 p=0.310
Upper GI/CNS Haemorrhage needing hospital admission	Number of upper GI/CNS Haemorrhage needing hospital admission Aspirin “No”	Number of upper GI/CNS Haemorrhage needing hospital admission Aspirin “Yes”	HR Unadjusted	HR Adjusted
Cohort total	31772	11275	N/A	N/A
Basic cox model analysis from Questionnaire 2 for aspirin “Yes” versus “No”	131	96	HR 1.64 95% 1.25-2.15 p=<0.0001	HR 1.21 95% CI 0.78-1.89 p=0.400

4.4 Discussion

The rationale for this study was as part of preliminary work to assess the feasibility of a proposed larger randomised study assessing aspirin as a primary prevention agent against cancer. The aim was two-fold, to establish if EHR could be used as a basis for such a study and also to add to the epidemiological evidence of aspirin in the primary prevention setting of cancer where the evidence is still debated. The study demonstrated that it was feasible to access certain data sets such HES/PEDW, ONS and cancer registries and obtain appropriate data on cancer incidence and mortality which would be important for any long-term aspirin prevention study. In terms of adding to the evidence base for the preventative effects of aspirin several of the analyses showed no effect as described above though there was evidence of an effect on specific tumour types in cohort A.

The key challenge for this study was the amount of missing data relating to when aspirin was started and stopped. This resulted in two cohorts being defined. The two different approaches to the analysis raise the question as to which of the analyses is more accurate. The analysis for cancer incidence that only includes cancers diagnosed after 2014 is proposed as the “cleaner” of the two analyses as the cancer definitely followed the participant reporting that they were taking aspirin. However the survival analysis from randomisation which uses imputed start and stop dates for aspirin tries to account for this missing data and also takes in to account the length of aspirin use which in previous studies has been shown to be significant for cancer prevention (189). Imputation models for missing data have previously been used both in epidemiological studies and RCTs and when used correctly they can assist analysis in this setting (201). This also has the benefit of more subjects and longer follow-up particularly in the relatively small cohorts of breast, colorectal and lung cancer studied in this analysis.

Finally, cohort B had a relatively smaller follow-up of 4 years in England, the majority of patients, and potentially 2 years from those who lived in Wales when compared to cohort A which could be over 10 years. From previous work this reduced follow-up could be highly significant in assessing aspirin effect on cancer incidence and mortality (**Figure 4.1)**. This may mean that both analyses provide useful information in this

setting and reflects that there may be more evidence for aspirin primary prevention for certain individual cancers rather than all cancer incidence particularly cancers of digestive system (178, 179, 181).

In the secondary analysis of upper GI/ CNS bleeds, which were the primary cause for hospital admission in HES data, there was an increased risk of upper GI or CNS haemorrhage needing hospital admission for those taking aspirin in keeping with previous literature. In the adjusted model for confounders the statistical significance was lost. There were relatively few events in this category which is also demonstrated in recent RCTs where the event rate for both events are less than 1% of the cohort (183). This study supports the well-established evidence that aspirin is associated with a small increase in the risk of bleeding.

Aspirin use as a primary preventative agent in cancer is still debated especially with a difference between findings between cohort studies and RCTs. Many feel that observational studies overstate the effect of aspirin as a preventative agent and argue that inappropriate statistical analysis can potentially cause bias towards a positive finding (202). Additionally, many of the RCTs were not initially set up for consideration of aspirin as chemopreventative agent against cancer but rather as cardiovascular disease preventative agent. Recent meta-analysis conflict previous publications and argue there is less evidence in RCT data to support aspirin's role as a primary preventative agent in cancer (186, 193). These are due primarily to recent studies such as the ASPREE, ASCEND, and ARRIVE. These have had criticism as they either do not have a long enough follow-up or possibly in too old an age group to have a significant effect on cancer (183, 185, 203).

There are therefore still significant questions about aspirin's role in cancer prevention. This study provides evidence to support long-term follow-up with EHR should be feasible and demonstrates the potential importance of developing further RCTs. One strategy would be to focus on individuals at a high risk of developing cancer to get the best benefit: risk ratio. For example in Lynch syndrome the CAPP2 trial (ISRCTN59521990) has shown a decreased risk of colorectal cancer with the use of aspirin and this has led to a change in practice and recommendations in NICE guidelines for the use of aspirin in those with Lynch syndrome (190, 204, 205).

This study using UKCTOCS data has limitations as discussed above the main challenge was missing data. The imputation model from recruitment and following questionnaire 2 was used to negate this as much as possible to achieve the most accurate and correct result with the data that was available. The missing data resulted from the fact that it was from an additional questionnaire added to the UKCTOCS trial during the follow-up period. The questionnaires which were vital for aspirin information were not specifically designed to answer the hypothesised question in this study. As shown with similar studies data at baseline, accurate wording of questions on dose, duration and compliance to aspirin use is extremely important to ascertain the effect of aspirin on cancer prevention (197, 198, 200).

This was also important for confounders selection which was based on information given by the participant via the questionnaires. The only confounder to be significant was statin use in favour of aspirin as a cancer preventative agent. This cannot be initially explained clinically but a hypothesis why this might be the case is that these drugs are taken in the events of CVS disease. CVS disease is often associated with platelet activation. Platelet activation is also involved in cancer development and spread and therefore may be associated with a decrease in cancer incidence.

Moving forward and considering the potential design of further cohort studies or a new RCT certain challenges have been identified. At present there is no one electronic database that can give prescription data for the whole population of the UK. In England GP data may be made available for research which may allow for more accurate recording of aspirin use (206). However, a questionnaire strategy or healthcare professional questionnaire to ascertain accurate information on aspirin use may still be needed in a pragmatic long-term trial. Another limitation to this study is that this is only a cohort in women so cannot be necessarily generalised to men as there are different cancer types.

This study does demonstrate, as with other evidence in this thesis, that cancer registries and potentially HES data can be used effectively to collect outcome data for a pragmatic trial with cancer incidence and mortality being at the present time the most readily available data. However as shown by the censor dates of the cancer registries there is still a significant lag in data which may impact the reporting times and costs of a future trial. There is now a proposed rapid cancer registry with a markedly decreased

lag time (167). However, there is a continued concern that EHR may not have the accuracy needed for a RCT. The UKCTOCS group have recently shown that cancer registration data may not be sufficient to get an accurate picture of cancer outcomes for RCTs and information from multiple sources such as treating physician and patient may also be needed (154). Data in chapter 2 of this thesis also supports this. Similarly, as shown in Chapter 1 serious adverse event data has the same limitations and should be taken into account when designing pragmatic RCTs in this setting and also in the writing of protocols.

4.5 Conclusion

This study demonstrates the ongoing equipoise of evidence for aspirin in the primary prevention setting and therefore the need for further RCTs. It also highlights some of the limitations of cohort studies. This study had challenges with missing data, however, care was taken to analyse the data in several ways to provide the most information for interpretation. Fundamentally this study was done as provisional work for a large pragmatic study in aspirin to see if EHR could be used in this setting. This adds to the evidence that EHR could be one strategy to follow-up a large cohort of participants in a RCT over a period of 10 years or more. Further work, following this thesis, will be done to devise a trial in chemotherapeutic primary prevention.

5 Thesis Conclusions

Oncology RCTs are one of the main drivers for innovation and change in practice within cancer care. They are particularly pertinent to the development and regulatory approval of new anti-cancer drugs. However, RCTs are increasingly expensive and cause a significant burden to healthcare systems (16, 17). This is particularly the case in the UK where resources for research can be limited (134). Participants within oncology trials also often need long-term follow-up due to the nature of the disease, and to monitor for any long-term side effects.

There are many potential ways to try to make clinical trials more cost effective and efficient (133). One way that has been considered is that in the future more trials could use EHR data i.e. data routinely collected in a healthcare system (49). There is continued questions, however, from researchers and regulators on the utility (reliability, completeness, accuracy) and accessibility of the data (25).

This thesis aimed to address some of those concerns and to find methods for how this data could be used in the future to improve the efficiency of established clinical trials and novel pragmatic cancer trials. I used trial and EHR data from three established trials managed by the MRC CTU at UCL to address specific questions related to the EHR data.

5.1 Main conclusions

In the PATCH trial analysis EHR data from HES APC and NICOR was compared with trial data from CRFs for the cardiovascular outcomes within the trial. This demonstrated that there is marked variability between these three sources of data if compared without clinical scrutiny. The EHR data missed important events where there was no inpatient hospital admission for example out of hospital silent MI, PE or TIA. HES APC data and NICOR data was sensitive for trial data when the event was an inpatient hospital admission, and the HF/ACS event was the main reason for admission at 0.89. However, the accuracy of diagnostic coding in HES APC data was variable based on whether the event was the primary diagnosis (primary reason for admission) or a secondary diagnosis. This raises the possibility that if the EHR were

solely used for this purpose a significant number of false positives would be identified. To use EHR within trials more widely in the future internationally recognised definitions of a trial SAEs will need to be developed as well as an implementation strategy.

The strength of the NICOR dataset is that there is enough clinical information to permit accurate diagnosis. The weakness is that it may miss some events that did not meet its strict diagnosis coding criteria but are still clinically significant for an individual patient within a trial. Trial data can also miss certain events especially if the participant has been in follow-up for a significant amount of time. A combination of data sources may provide the broadest and most accurate method of capturing these events from multiple healthcare settings.

The Add-Aspirin data also adds to this debate demonstrating that there was poor concordance between HES APC data and Add-Aspirin CRF data for protocol defined adverse events with possible high levels of false positives within the EHR data. Other studies in different disease areas have shown EHR either miss or over report clinical events as demonstrated in the Standard and New Antiepileptic Drugs II trial (SANAD-II (ISRCTN30294119)) where EHR underestimated seizure events compared to traditional CRF reported events (83, 207).

The lesson for future trials is that multiple sources of information might be best to get the most accurate information for a patient's adverse events and clinical scrutiny may also be needed. The trial must therefore from the start define what an event consists of based not only on clinical definition but also ICD-10 codes and be transparent in the protocol and also on publication how the events were defined.

The length of time it took to access data from NICOR and NHS Digital for this project is a concern. Our work, documented through publication, demonstrates that when you have not formally planned for EHR access at the start of the trial it is difficult to obtain (120). The access to both NICOR and NHS Digital data for the PATCH methodological study took years due to a change in data laws and also strict consent wording. Even with the Add-Aspirin application, where preliminary discussions with the registry were had, the length of time to access and transfer data took 13 months after the first application was submitted. Lastly the retention of the data is also a significant issue with some of the registries not allowing extended contracts meaning renewal policies annually which then costs hundreds of pounds to keep the data or even over a

thousand if an amendment is needed. For a trial this is not practical in loss of time spent filling in forms and financially as, depending on the study, you may need to keep the primary data for a minimum of 10 years following the closure of the trial.

The Add-Aspirin trial data was compared with datasets that could be accessed, at the time of writing, held by NCRAS and applied for via ODR. This demonstrated that death data and also cancer registration, including basic staging, is of a high standard for those cancers being assessed in this trial (breast, prostate, colorectal and gastro-oesophageal). The data also showed that recurrence data has improved from previously documented analyses (75, 151, 155). However, it does require access to all cancer registries as well as HES APC data to see comparable rates of recurrence compared to trial data. The method explored in this piece of work is also dependent on knowing when radical treatment had been completed. I concluded that the cancer registry data could be a very good resource as an adjunct to trial data for the primary outcome of recurrence within the Add-Aspirin trial. This would allow for greater confidence in the primary outcome analysis and also to reduce the amount of work required by sites in following up these patients. This could be an invaluable resource for other oncology trials in the future.

There are some limitations to the use of EHR data such as the time lag of up to 18 months to receive cancer registration data This could be extremely difficult for trials where they need to complete their data collection within a certain time frame. The prolongation of a trial can be significant for funders, sponsors and regulators of drugs. The cancer registry can provide rapid access to cancer registration data however this early data might not be of significant detail or reliability to use in cancer trials (167).

We did also see some delays in the trial data collection. This is due to the nature of collection of some data as it is usually only collected following a defined trial visit with a gap of possibly as long as a year between visits in long-term follow-up. For some of the events such as death the outcome was collected quicker and to the same standard in EHR data compared to trial data. This increases the evidence that some outcomes particularly death could just be gathered from EHR sources to have the most efficient trial data collection.

Finally, the last chapter used UKCTOCS data as preliminary work for a larger RCT in primary prevention of cancer using aspirin. This helped define how EHR data could be

used in a guide for protocol development and data sharing applications in the future trial. However, data from the UKCTOCS cohort of around 50,000 women did not demonstrate a clear benefit of aspirin in decreasing cancer incidence or overall cancer mortality. This could be due to missing aspirin data within the trial as the length of time and compliance of medication was not robust. There was some evidence for aspirin use decreasing the risk for individual cancers including breast and colorectal which would agree with previous epidemiological studies. Aspirin use as a primary cancer prevention agent is still under debate and therefore further pragmatic RCTs in this setting are crucial to try to answer this important question. Lastly this piece of work demonstrated the importance of questionnaire question design and how questionnaires can be used to good effect to collect trial data but possibly only as supplementation to other more objective sources.

5.2 Future challenges for EHR in clinical trials and solutions

There is great demand for EHR use within clinical trials (64). This thesis demonstrates that EHRs have great potential in many ways to improve trial data collection and also possibly improve the financial burden and logistical burden for trial sites/ clinical trial units. However, there are still major hurdles that need to be overcome.

It is challenging to find out which trials have used EHR as publications don't always mention the source of data and traditional systematic review searches are often ineffective in this setting. However in a systematic review of all trials which registered for data access there is still only a minority of UK clinical trials using EHR in their analyses (33). For trialists to be confident in using EHR resources access to data also has to be improved. Data applications for academic clinical trials must become simpler or more guidance through the application process is needed from data holders. There needs to be one point of access for all the devolved nations as at present multiple applications are needed for each country. Datasets also need to be held by fewer data controllers to decrease the cost and time needed to fill in data sharing applications.

If an institution complies with verified standards of data storage and governance, then this aspect of the application could be condensed and streamlined for those

applicants. Data holders also need to work with trialists to establish contracts that are financially and logistically feasible to hold and update the data for long periods of time. EHR data needs to be able to be stored by clinical trial units in a similar way to trial data for long-term use. The primary data from academic work must also be available for scrutiny. However, at present it is extremely difficult to share this data easily with another institution as part of the data sharing contracts with data holders. The sharing of data with other academic institutions/ regulators needs to be clarified and made possible. Clinical trials units need to work with regulators to allow EHR data to be considered a primary data source so that it can be used in applications for licensing medication.

There is ongoing work in all these areas with the establishment of NHS Digitrials who are working with trialists to make applications and EHR data appropriate for clinical trials use. Health Data Research (HDR) UK are forming collaborations and central resources of information about EHR registries in an attempt to demystify the application process (208). They have also commissioned a cross-programme call for 'Data enabled trials' with 7 future trials funded following the call (209). NHS Digital are continuing to try and manage/ hold more of the English datasets to reduce the amount of applications. An example of this is the integration of the cancer registry to NHS Digital and also the creation of a GP data resource which will cover the whole of England (206). MHRA are also working to make EHR data a primary resource that can be used within trials by giving guidance on how to audit the data (88). There is ongoing work within the MRC CTU with NHS Digital to demonstrate that HES data can be audited to equivalent MHRA standard to make sure that is complete, consistent, enduring and available.

The COVID pandemic has shown that a lot of these areas of concern can be improved with financial aid and political will. This can be seen with the development CVD-COVID-UK Consortium which was allowed unprecedented access to deidentified data in England via NHS Digital Trusted Research Environment during the COVID crisis (123). This was more for epidemiological studies but COVID trials were also were able to gather routine data quickly to help assist data collection as seen with the RECOVERY trial (NCT04381936) (210). This speed of access should be maintained for future trials and studies in the future now that the infrastructure has been set up.

The work in this thesis demonstrates that in the future trialists need to have early discussions with data holders to establish appropriate trial protocols and how any EHR data will be used to assess outcomes. Trialists should have access to dummy datasets that simulate datasets for example Simulacrum for NCRAS data. This would allow researchers to create appropriate coding for extraction and develop protocols (211). Trialists need to use recently developed consort definitions to design and publish their work (127). Finally new trials need to be designed along with methodologists to establish how best to use EHR so that it achieves reliable and accurate outcomes for trials. This should also include cost effectiveness studies to run along side these trials to make sure that EHR does actually provide a cost benefit overall to funders.

This thesis has been focused on UK trials but to improve international use of EHR then other countries data must be improved through international collaborations of data definitions and standards. This is already established in many counties in Europe and America but to ensure more diverse clinical trials in lower economic countries then this must be a wider consideration for policy makers. We have seen the variability of data from across the world during the COVID crisis and how influential this can be on recognising how health interventions effect populations on a global scale. This would be an invaluable resource to help global collaboration in trials.

Table 5-1 Key consideration for EHR use in trials

Key Considerations for EHR use in trials
1. If EHR is to be used for trial outcome collection then this should be decided at the start of the trial. Ideally the trial outcomes should be defined to match the appropriate coding of the EHR dataset and also the clinical setting that the EHR is collecting data from. 2. When using EHR datasets with clinical trials then the dataset must be auditable to trace the source and validity of the data. Work should also be done to assess the validity of the EHR data with the appropriate trial outcome with clinical trial methodology research. 3. EHR dataset accessibility is difficult due to strict data governance laws. This will hopefully improve with development of trusted research environments. Before attempting to access the data, it is important a trial team has an understanding of these trusted research environments and appropriate knowledge in how to work with the individual EHR datasets. 4. Researchers need to continue to work with data controllers of EHR to make data accessible and in a format that can be used for research. Also national datasets need to be easily linked for researchers and, in the future, devolved nations' datasets should be accessed from one data controller to save time and money. 5. Publication of trials using EHR need to follow consort guidelines to make sure that the appropriate information for readers is clearly stated in the paper and the methodology is transparent and reproducible.

5.3 Future research

Following on from this work I am now working in collaboratives to try and set up a primary prevention study for cancer and potentially other important health outcomes of ageing integrating EHR into the protocol and follow-up. I hope to use the skills that I have learned in this thesis to help design and implement these studies with novel designs. I think it is important to continue to assess which outcomes are possible to collect using EHR data and what could be possible in the future. This is an evolving process as the EHR/ registries continue to develop and add new datasets become available. I would also like to continue to gain knowledge of other devolved nations national registries which were out of scope of this thesis to create truly nationwide pragmatic RCTs. Future areas of integration could be from electronic PROMS and patient questionnaires to run along side EHR. This would not only give more emphasis to the patients' perception of the outcome but help to develop more effective ways to do trials with less finance and administrative staffing needed. EHR continues to improve all the time but we must make sure, as with all evolving technologies, that we analyse them to make sure they enhance trials rather than disrupt.

6 References

1. Mukerjee S. The Emperor of all Maladies: A biography of cancer. New York: Scribner; 2010.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021; 71: 209- 249. doi: 10.3322/caac.21660.
3. Cancer research UK. Cancer incidence Statistics. [Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence. Accessed 20/02/2020.
4. NHS England. Achieving World Class Cancer outcomes: taking the stratey foward 2016. [Available from: https://www.england.nhs.uk/wp-content/uploads/2016/05/cancer-strategy.pdf. Accessed 02/03/2020.
5. NHS England. Next steps on the NHS five year forward view 2017. [Available from: https://www.england.nhs.uk/wp-content/uploads/2017/03/NEXT-STEPS-ON-THE-NHS-FIVE-YEAR-FORWARD-VIEW.pdf. Accessed 20/02/2020.
6. Cancer Research UK. Cancer diagnosis and treatment statistics 2017. [Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/diagnosis-and-treatment#heading-Five. Accessed 20/02/2020.
7. Thompson MK, Poortmans P, Chalmers AJ, Faivre-Finn C, Hall E, Huddart RA, et al. Practice-changing radiation therapy trials for the treatment of cancer: where are we 150 years after the birth of Marie Curie? British Journal of Cancer. 2018;119(4):389-407. doi: 10.1038/s41416-018-0201-z.
8. Kendall JM. Designing a research project: randomised controlled trials and their principles. Emerg Med J. 2003 Mar;20(2):164-8. doi: 10.1136/emj.20.2.164.
9. Sibbald B, Roland M. Understanding controlled trials: Why are randomised controlled trials important? BMJ. 1998;316(7126):201. doi: 10.1136/bmj.316.7126.201.
10. James S, Rao SV, Granger CB. Registry-based randomized clinical trials—a new clinical trial paradigm. Nature Reviews Cardiology. 2015;12:312. doi: 10.1038/nrcardio.2015.33.
11. Savage P, Mahmoud S. Development and economic trends in cancer therapeutic drugs: a 5-year update 2010-2014. British journal of cancer. 2015;112(6):1037-41. doi: 10.1038/bjc.2015.56.
12. Mullard A. 2018 FDA drug approvals. Nature Reviews Drug Discovery. 2019;18:85-9. doi: 10.1038/d41573-019-00014-x.
13. Bothwell LE, Greene JA, Podolsky SH, Jones DS. Assessing the Gold Standard — Lessons from the History of RCTs. 2016;374(22):2175-81. doi: 10.1056/NEJMms1604593.
14. Cancer research UK. We're saving lives through research 2019. [Available from: https://www.cancerresearchuk.org/sites/default/files/ec1060588_cruk_ar_2019_interactive.pdf. Accessed 20/02/2020.
15. NIHR. Our studies 2019. [Available from: https://www.nihr.ac.uk/explore-nihr/specialties/cancer.htm. Accessed 20/02/2020.
16. Speich B, von Niederhäusern B, Blum CA, Keiser J, Schur N, Fürst T, et al. Retrospective assessment of resource use and costs in two investigator-initiated

randomized trials exemplified a comprehensive cost item list. Journal of Clinical Epidemiology. 2018;96:73-83. doi: 10.1016/j.jclinepi.2017.12.022.
17. Sertkaya A, Wong H-H, Jessup A, Beleche T. Key cost drivers of pharmaceutical clinical trials in the United States. Clin Trials. 2016;13(2):117-26. doi: 10.1177/1740774515625964.
18. Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JPA. Routinely collected data and comparative effectiveness evidence: promises and limitations. CMAJ. 2016;188(8):E158-E64. doi: 10.1503/cmaj.150653.
19. Rothwell PM. External validity of randomised controlled trials: "To whom do the results of this trial apply?". The Lancet. 2005;365(9453):82-93. doi: 10.1016/S0140-6736(04)17670-8.
20. Rothwell PM. Commentary: External validity of results of randomized trials: disentangling a complex concept. International Journal of Epidemiology. 2010;39(1):94-6. doi: 10.1093/ije/dyp305.
21. Davies G, Jordan S, Brooks CJ, Thayer D, Storey M, Morgan G, et al. Long term extension of a randomised controlled trial of probiotics using electronic health records. Scientific Reports. 2018;8(1):7668. doi:10.1038/s41598-018-25954-z.
22. Leak C, Goggins K, Schildcrout JS, Theobald C, Donato KM, Bell SP, et al. Effect of Health Literacy on Research Follow-Up. J Health Commun. 2015;20 Suppl 2(0):83-91. doi:10.1080/10810730.2015.1058442.
23. Woolard RH, Carty K, Wirtz P, Longabaugh R, Nirenberg TD, Minugh PA, et al. Research Fundamentals: Follow-up of Subjects in Clinical Trials: Addressing Subject Attrition. Academic Emergency Medicine. 2004;11(8):859-66. doi: 10.1111/j.1553-2712.2004.tb00769.x.
24. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, et al. Electronic health records to facilitate clinical research. Clinical research in cardiology : official journal of the German Cardiac Society. 2017;106(1):1-9. doi: 10.1007/s00392-016-1025-6.
25. McCord K, Hemkens L. Using electronic health records for clinical trials: Where do we stand and where can we go? Canadian Medical Association Journal. 2019;191(5):E128-E33. doi: 10.1503/cmaj.180841.
26. Bereznicki BJ, Peterson GM, Jackson SL, Walters EH, Fitzmaurice KD, Gee PR. Data-mining of medication records to improve asthma management. Medical Journal of Australia. 2008;189(1):21-5. doi: 10.5694/j.1326-5377.2008.tb01889.x.
27. Woodcock A, Bakerly ND, New JP, Gibson JM, Wu W, Vestbo J, et al. The Salford Lung Study protocol: a pragmatic, randomised phase III real-world effectiveness trial in asthma. BMC Pulmonary Medicine. 2015;15(1):160. doi: 10.1186/s12890-015-0150-8.
28. Tamblyn R, Ernst P, Winslade N, Huang A, Grad R, Platt RW, et al. Evaluating the impact of an integrated computer-based decision support with person-centered analytics for the management of asthma in primary care: a randomized controlled trial. JAMIA. 2015;22(4):773-83. doi: 10.1093/jamia/ocu009.
29. Price M, Davies I, Rusk R, Lesperance M, Weber J. Applying STOPP Guidelines in Primary Care Through Electronic Medical Record Decision Support: Randomized Control Trial Highlighting the Importance of Data Quality. JMIR Med Inform. 2017;5(2):e15-e. doi: 10.2196/medinform.6226.
30. World Health Organisation. World health statistics 2019: monitoring health for the SDGs, sustainable development goals. Geneva: World Health Organization; 2019. [Avialable from :

https://apps.who.int/iris/bitstream/handle/10665/324835/9789241565707-eng.pdf?sequence=9&isAllowed=y
31. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394-424. doi: 10.3322/caac.21492.
32. Ferlay J, Colombet M, Soerjomataram I, Dyba T, Randi G, Bettio M, et al. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. European Journal of Cancer. 2018;103:356-87. doi: 10.1016/j.ejca.2018.07.005.
33. Lensen S, Macnair A, Love SB, Yorke-Edwards V, Noor NM, Martyn M, et al. Access to routinely collected health data for clinical trials – review of successful data requests to UK registries. Trials. 2020;21(1):398. doi:10.1186/s13063-020-04329-8.
34. Chief Data Officer. NHS hospital data and Datasets: A consultation. In: NHS Digital, editor. 2013. [Available from: https://www.england.nhs.uk/wp-content/uploads/2013/07/hosp-data-consult.pdf
35. Modernising health and care public bodies- the health and social care act 2012. 2012.
36. Boyd A, . Understanding Hospital Episode Statistics (HES). London, UK: CLOSER,. 2017. [Available from: https://www.closer.ac.uk/wp-content/uploads/CLOSER-resource-Understanding-HES.pdf
37. Thorn JC, Turner E, Hounsome L, Walsh E, Donovan JL, Verne J, et al. Validation of the Hospital Episode Statistics Outpatient Dataset in England. Pharmacoeconomics. 2016;34(2):161-8. doi:10.1007/s40273-015-0326-3.
38. Clements C, Turnbull P, Hawton K, Geulayov G, Waters K, Ness J, et al. Rates of self-harm presenting to general hospitals: a comparison of data from the Multicentre Study of Self-Harm in England and Hospital Episode Statistics. BMJ Open. 2016;6(2):e009749. doi: 10.1136/bmjopen-2015-009749.
39. NHS Digital. Data Access Request Service (DARS) products and services 2019. [Available from: https://digital.nhs.uk/services/data-access-request-service-dars/dars-products-and-services. Accessed 09/08/2019.
40. NHS Digital. Data Access Request Service (DARS): process 2019. [Available from: https://digital.nhs.uk/services/data-access-request-service-dars/data-access-request-service-dars-process. Accessed 20/02/2020.
41. NHS Digital. NHS DigiTrials 2020. [Available from: https://digital.nhs.uk/services/nhs-digitrials. Accessed 20/02/2020.
42. Parkin DM. The evolution of the population-based cancer registry. Nature Reviews Cancer. 2006;6(8):603-12. doi: 10.1038/nrc1948.
43. NHS Digital. Ukiacr. [Available from: http://www.ukiacr.org/. Accessed 09/08/2019.
44. Public Health England. National Cancer Intelligence Network (NCIN): 30 + years of cancer intelligence- challenges of technologies of the time. [Available from: http://www.ncin.org.uk/home. Accessed 09/09/2019.
45. Appleyard SE, Gilbert DC. Innovative Solutions for Clinical Trial Follow-up: Adding Value from Nationally Held UK Data. Clinical Oncology. 2017;29(12):789-95. doi: 10.1016/j.clon.2017.10.003.
46. Healthcare Quality Improvment Partnership. The National Clinical Audit Programme 2019. [Available from: https://www.hqip.org.uk/a-z-of-nca/#.XU1WfuRYboo. Accessed 09/08/2019.

47. NICOR. NICOR 2020. [Available from: https://www.nicor.org.uk/. Accessed 20/02/2020.
48. Datalink CPR. CPRD UK data driving real world evidence 2019. [Available from: https://cprd.com/home. Accessed 09/08/2019.
49. Lauer MS, D'Agostino RB. The Randomized Registry Trial — The Next Disruptive Technology in Clinical Research? New England Journal of Medicine. 2013;369(17):1579-81. doi: 10.1056/NEJMp1310102.
50. Li G, Sajobi TT, Menon BK, Korngut L, Lowerison M, James M, et al. Registry-based randomized controlled trials- what are the advantages, challenges, and areas for future research? Journal of Clinical Epidemiology. 2016;80:16-24. doi: 10.1016/j.jclinepi.2016.08.003.
51. Liu JB, D'Angelica MI, Ko CY. The Randomized Registry Trial: Two Birds, One Stone. Ann Surg. 2017;265(6):1064-5. doi: 10.1097/SLA.0000000000002166.
52. Lagerqvist B, Fröbert O, Olivecrona GK, Gudnason T, Maeng M, Alström P, et al. Outcomes 1 Year after Thrombus Aspiration for Myocardial Infarction. New England Journal of Medicine. 2014;371(12):1111-20. doi: 10.1056/NEJMoa1405707.
53. Reilly R, Paranjothy S, Beer H, Brooks C, Fielder H, Lyons R. Birth outcomes following treatment for precancerous changes to the cervix: a population-based record linkage study. BJOG. 2012;119(2):236-44. doi: 10.1111/j.1471-0528.2011.03052.x.
54. Cairns V, Wallenhorst C, Rietbrock S, Martinez C. Incidence of Lyme disease in the UK: a population-based cohort study. BMJ Open. 2019;9(7):e025916. doi: 10.1136/bmjopen-2018-025916.
55. Herrett E, Gadd S, Jackson R, Bhaskaran K, Williamson E, van Staa T, et al. Eligibility and subsequent burden of cardiovascular disease of four strategies for blood pressure-lowering treatment: a retrospective cohort study. The Lancet. 2019;394(10199):663-671. doi: 10.1016/S0140-6736(19)31359-5
56. Filion KB, Azoulay L, Platt RW, Dahl M, Dormuth CR, Clemens KK, et al. A Multicenter Observational Study of Incretin-based Drugs and Heart Failure. New England Journal of Medicine. 2016;374(12):1145-54. doi: 10.1056/nejmoa1506115.
57. Yska JP, van Roon EN, de Boer A, Leufkens HGM, Wilffert B, de Heide LJM, et al. Remission of Type 2 Diabetes Mellitus in Patients After Different Types of Bariatric Surgery: A Population-Based Cohort Study in the United Kingdom. JAMA Surgery. 2015;150(12):1126-33. doi:10.1001/jamasurg.2015.2398.
58. Hemkens LG. How Routinely Collected Data for Randomized Trials Provide Long-term Randomized Real-World Evidence.JAMA Network Open. 2018;1(8):e186014. doi:10.1001/jamanetworkopen.2018.6014.
59. Coleman R, Cameron D, Dodwell D, Bell R, Wilson C, Rathbone E, et al. Adjuvant zoledronic acid in patients with early breast cancer: final efficacy analysis of the AZURE (BIG 01/04) randomised open-label phase 3 trial. The Lancet Oncology. 2014;15(9):997-1006. doi: 10.1016/S1470-2045(14)70302-X.
60. Jacobs IJ, Menon U, Ryan A, Gentry-Maharaj A, Burnell M, Kalsi JK, et al. Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. The Lancet. 2016;387(10022):945-56. doi: 10.1016/S0140-6736(15)01224-6.
61. Ford I, Murray H, Packard CJ, Shepherd J, Macfarlane PW, Cobbe SM. Long-Term Follow-up of the West of Scotland Coronary Prevention Study. 2007;357(15):1477-86. doi: 10.1056/NEJMoa065994.

62. The West of Scotland Coronary Prevention Study Group. Computerised record linkage: Compared with traditional patient follow-up methods in clinical trials and illustrated in a prospective epidemiological study. Journal of Clinical Epidemiology. 1995;48(12):1441-52. doi: 10.1016/0895-4356(95)00530-7.
63. Cottrell DJ, Wright-Hughes A, Collinson M, Boston P, Eisler I, Fortune S, et al. Effectiveness of systemic family therapy versus treatment as usual for young people after self-harm: a pragmatic, phase 3, multicentre, randomised controlled trial. The Lancet Psychiatry. 2018;5(3):203-16. doi: 10.1016/S2215-0366(18)30058-0.
64. McKay AJ, Jones AP, Gamble CL et al. Use of routinely collected data in a UK cohort of publicly funded randomised clinical trials. F1000Research. 2020;9:323. doi:10.12688/f1000research.23316.2.
65. NHS Digital. Data Access Request Service (DARS) charges 2018/19 2019. [Available from: https://digital.nhs.uk/services/data-access-request-service-dars/data-access-request-service-dars-charges-2018-19. Accessed 09/08/2019.
66. Morris EJA, Jordan C, Thomas JD, Cooper M, Brown JM, Thorpe H, et al. Comparison of treatment and outcome information between a clinical trial and the National Cancer Data Repository. Br J Surg. 2011;98(2):299-307. doi: 10.1002/bjs.7295.
67. Merriel SWD, Turner EL, Walsh E, Young G, Metcalfe C, Hounsome L, et al. Validation of the National Cancer Registration and Analysis Service prostate cancer registry with data from the CAP study. The Lancet. 2016;388:S77. doi: 10.1016/S0140-6736(16)32313-3.
68. Larsen IK, Småstuen M, Johannesen TB, Langmark F, Parkin DM, Bray F, et al. Data quality at the Cancer Registry of Norway: An overview of comparability, completeness, validity and timeliness. European Journal of Cancer. 2009;45(7):1218-31. doi: 10.1016/j.ejca.2008.10.037.
69. Gentry-Maharaj A, Fourkala EO, Burnell M, Ryan A, Apostolidou S, Habib M, et al. Concordance of National Cancer Registration with self-reported breast, bowel and lung cancer in England and Wales: a prospective cohort study within the UK Collaborative Trial of Ovarian Cancer Screening. British journal of cancer. 2013;109(11):2875-9. doi: 10.1038/bjc.2013.626.
70. Korhonen P, Malila N, Pukkala E, Teppo L, Albanes D, Virtamo J. The Finnish Cancer Registry as Follow-Up Source of a Large Trial Cohort. Acta Oncologica. 2002;41(4):381-8. doi: 10.1080/028418602760169442.
71. Nilbert M, Thomsen LA, Winther Jensen J, Møller H, Borre M, Widenlou Nordmark A, et al. The power of empirical data; lessons from the clinical registry initiatives in Scandinavian cancer care. Acta Oncol. 2020;59(11):1343-56. doi: 10.1080/0284186X.2020.1820573.
72. Zanetti R, Schmidtmann I, Sacchetto L, Binder-Foucard F, Bordoni A, Coza D, et al. Completeness and timeliness: Cancer registries could/should improve their performance. European Journal of Cancer. 2015;51(9):1091-8. doi: 10.1016/j.ejca.2013.11.040.
73. Parkin DM, Bray F. Evaluation of data quality in the cancer registry: Principles and methods Part II. Completeness. European Journal of Cancer. 2009;45(5):756-64. doi: 10.1016/j.ejca.2008.11.033.
74. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: Principles and methods. Part I: Comparability, validity and timeliness. European Journal of Cancer. 2009;45(5):747-55. doi: 10.1016/j.ejca.2008.11.032.

75. Henson KE, Elliss-Brookes L, Coupland VH, Payne E, Vernon S, Rous B, et al. Data Resource Profile: National Cancer Registration Dataset in England. International Journal of Epidemiology. 2020;49(1):16-16h. doi: 10.1093/ije/dyz076.
76. NHS Digital. The processing cycle and HES data quality 2020. [Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/the-processing-cycle-and-hes-data-quality. Accessed 19/11/2020.
77. Brocklehurst P, Field D, Greene K, Juszczak E, Keith R, Kenyon S, et al. Computerised interpretation of fetal heart rate during labour (INFANT): a randomised controlled trial. The Lancet. 2017;389(10080):1719-29. doi: 10.1016/S0140-6736(17)30568-8.
78. Hagger-Johnson G, Harron K, Fleming T, Gilbert R, Goldstein H, Landy R, et al. Data linkage errors in hospital administrative data when applying a pseudonymisation algorithm to paediatric intensive care records. BMJ Open. 2015;5(8):e008118. doi: 10.1136/bmjopen-2015-008118.
79. Herrett E, Shah AD, Boggon R, Denaxas S, Smeeth L, van Staa T, et al. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. BMJ. 2013;346:f2350. doi: 10.1136/bmj.f2350.
80. Buyse M, Squifflet P, Coart E, Quinaux E, Punt CJ, Saad ED. The impact of data errors on the outcome of randomized clinical trials. Clin Trials. 2017;14(5):499-506. doi: 10.1177/1740774517716158.
81. Data protection act 2018- factshett- overview. In: Department for Digital C, Media and Sport,. editor. 2018.
82. Partridge N. Review of data releases by the NHS Information Centre. 2014.
83. Powell GA, Bonnett LJ, Tudur-Smith C, Hughes DA, Williamson PR, Marson AG. Using routinely recorded data in the UK to assess outcomes in a randomised controlled trial: The Trials of Access. Trials. 2017;18(1):389. doi: 10.1186/s13063-017-2135-9.
84. Lugg-Widger FV AL, Cannings-John R, Hood K, Hughes K, Moody G, Robling M. Challenges in accessing routinely collected data from multiple providers in the UK for primary studies: managing the morass. International Journal of Population Data Science. 2018;3(3). doi:10.23889/ijpds.v3i3.432.
85. Peden CJ, Stephens T, Martin G, Kahan BC, Thomson A, Rivett K, et al. Effectiveness of a national quality improvement programme to improve survival after emergency abdominal surgery (EPOCH): a stepped-wedge cluster-randomised trial. The Lancet. 2019;393(10187):2213-21. doi: 10.1016/S0140-6736(18)32521-2.
86. Drazen JM. Sharing Individual Patient Data from Clinical Trials. New England Journal of Medicine. 2015;372(3):201-2. doi: 10.1056/NEJMp1415160.
87. Medicines and Healthcare products Regulatory Agency. Good Clinical Practice Guide: TSO; 2012.
88. Medicines and Healthcare products Regulatory Agency. MHRA draft guidance on randomised controlled trials generating real-world evidence to support regulatory decisions 2020. [Available from: https://www.gov.uk/government/consultations/mhra-draft-guidance-on-randomised-controlled-trials-generating-real-world-evidence-to-support-regulatory-decisions. Accessed 20/11/2020.
89. Phillips R, Hazell L, Sauzet O, Cornelius V. Analysis and reporting of adverse events in randomised controlled trials: a review. BMJ Open. 2019;9(2):e024537. doi: 10.1136/bmjopen-2018-024537.

90. Péron J, Maillet D, Gan HK, Chen EX, You B. Adherence to CONSORT Adverse Event Reporting Guidelines in Randomized Clinical Trials Evaluating Systemic Cancer Therapy: A Systematic Review. Journal of Clinical Oncology. 2013;31(31):3957-63. doi: 10.1200/JCO.2013.49.3981.
91. Sivendran S, Latif A, McBride RB, Stensland KD, Wisnivesky J, Haines L, et al. Adverse Event Reporting in Cancer Clinical Trial Publications. Journal of Clinical Oncology. 2014;32(2):83-9. doi: 10.1200/JCO.2013.52.2219.
92. Seruga B, Templeton AJ, Badillo FEV, Ocana A, Amir E, Tannock IF. Under-reporting of harm in clinical trials. The Lancet Oncology. 2016;17(5):e209-e19. doi: 10.1016/S1470-2045(16)00152-2.
93. Edwards BJ, Gounder M, McKoy JM, Boyd I, Farrugia M, Migliorati C, et al. Pharmacovigilance and reporting oversight in US FDA fast-track process: bisphosphonates and osteonecrosis of the jaw. The Lancet Oncology. 2008;9(12):1166-72. doi: 10.1016/S1470-2045(08)70305-X.
94. Bishopric NH, Lippman ME. Adverse Cardiovascular Events in Cancer Trials. Missing in Action? Journal of the American College of Cardiology. 2020;75(6):629-31. doi: 10.1016/j.jacc.2019.12.019.
95. Bonsu JM, Guha A, Charles L, Yildiz VO, Wei L, Baker B, et al. Reporting of Cardiovascular Events in Clinical Trials Supporting FDA Approval of Contemporary Cancer Therapies. Journal of the American College of Cardiology. 2020;75(6):620-8. doi: 10.1016/j.jacc.2019.11.059.
96. Kivimäki M, Batty GD, Singh-Manoux A, Britton A, Brunner EJ, Shipley MJ. Validity of Cardiovascular Disease Event Ascertainment Using Linkage to UK Hospital Records. Epidemiology. 2017;28(5):735-9. doi: 10.1097/EDE.0000000000000688.
97. Rodrigues C, Odutayo A, Patel S, Agarwal A, Roza da Costa B, Lin E, et al. Comparison between cardiovascular trial outcomes and treatment effects using clinical endpoint committee adjudication versus routine health data: A systemic review. Canadian Journal of Cardiology. 2019;35(10):S13-S4. doi: 10.1161/CIRCOUTCOMES.120.007903
98. Barry SJ, Dinnett E, Kean S, Gaw A, Ford I. Are routinely collected NHS administrative records suitable for endpoint identification in clinical trials? Evidence from the West of Scotland Coronary Prevention Study. PLoS One. 2013;8(9):e75379. doi: 10.1371/journal.pone.0075379.
99. Gilbert DC, Duong T, Kynaston HG, Alhasso AA, Cafferty FH, Rosen SD, et al. Quality-of-life outcomes from the Prostate Adenocarcinoma: TransCutaneous Hormones (PATCH) trial evaluating luteinising hormone-releasing hormone agonists versus transdermal oestradiol for androgen suppression in advanced prostate cancer. BJU International. 2017;119(5):667-75. doi: 10.1111/bju.13687.
100. Culig Z, Santer FR. Androgen receptor signaling in prostate cancer. Cancer and Metastasis Reviews. 2014;33(2):413-27. doi: 10.1007/s10555-013-9474-0.
101. Huggins C, Scott WW, Hodges CV. Studies on Prostatic Cancer. III. The Effects of Fever, of Desoxycorticosterone and of Estrogen on Clinical Patients with Metastatic Carcinoma of the Prostate. Journal of Urology. 1941;46(5):997-1006.
102. Hayes FJ, Seminara SB, DeCruz S, Boepple PA, Crowley WF, Jr. Aromatase Inhibition in the Human Male Reveals a Hypothalamic Site of Estrogen Feedback. The Journal of Clinical Endocrinology & Metabolism. 2000;85(9):3027-35. doi: 10.1210/jcem.85.9.6795.
103. Finkelstein JS OL, Whitcomb RW, Crowley WF Jr. Sex Steroid Control of Gonadotropin Secretion in the Human Male. II. Effects of Estradiol Administration in

Normal and Gonadotropin-Releasing Hormone-Deficient Men. The Journal of Clinical Endocrinology & Metabolism. 1991;73(3):621-8. doi: 10.1210/jcem-73-3-621.
104. Wilkins A, Shahidi M, Parker C, Gunapala R, Thomas K, Huddart R, et al. Diethylstilbestrol in castration-resistant prostate cancer. BJU International. 2012;110(11b):E727-E35. doi: 10.1111/j.1464-410X.2012.11546.x.
105. Byar DP. Proceedings: The Veterans Administration Cooperative Urological Research Group's studies of cancer of the prostate. Cancer. 1973;32(5):1126-30.
106. Smith MR, Finkelstein JS, McGovern FJ, Zietman AL, Fallon MA, Schoenfeld DA, et al. Changes in Body Composition during Androgen Deprivation Therapy for Prostate Cancer. The Journal of Clinical Endocrinology & Metabolism. 2002;87(2):599-603. doi: 10.1210/jcem.87.2.8299.
107. Shahinian VB, Kuo Y-F, Freeman JL, Goodwin JS. Risk of Fracture after Androgen Deprivation for Prostate Cancer. New England Journal of Medicine. 2005;352(2):154-64. doi: 10.1056/NEJMoa041943.
108. Bosco C, Bosnyak Z, Malmberg A, Adolfsson J, Keating NL, Van Hemelrijck M. Quantifying observational evidence for risk of fatal and nonfatal cardiovascular disease following androgen deprivation therapy for prostate cancer: a meta-analysis. European urology. 2015;68(3):386-96. doi: 10.1016/j.eururo.2014.11.039.
109. O'Farrell S, Garmo H, Holmberg L, Adolfsson J, Stattin P, Van Hemelrijck M. Risk and timing of cardiovascular disease after androgen-deprivation therapy in men with prostate cancer. Journal of clinical oncology. 2015;33(11):1243-51. doi: 10.1200/JCO.2014.59.1792.
110. Sturgeon KM, Deng L, Bluethmann SM, Zhou S, Trifiletti DM, Jiang C, et al. A population-based study of cardiovascular disease mortality risk in US cancer patients. European Heart Journal. 2019;40(48):3889-97. doi: 10.1093/eurheartj/ehz766.
111. von Schoultz B, Carlström K, Collste L, Eriksson A, Henriksson P, Pousette Å, et al. Estrogen therapy and liver function—metabolic effects of oral and parenteral administration. The Prostate. 1989;14(4):389-95.
112. Langley RE, Godsland IF, Kynaston H, Clarke NW, Rosen SD, Morgan RC, et al. Early hormonal data from a multicentre phase II trial using transdermal oestrogen patches as first-line hormonal therapy in patients with locally advanced or metastatic prostate cancer. BJU international. 2008;102(4):442-5. doi: 10.1111/j.1464-410X.2008.07583.x
113. Langley RE, Kynaston HG, Alhasso AA, Duong T, Paez EM, Jovic G, et al. A Randomised Comparison Evaluating Changes in Bone Mineral Density in Advanced Prostate Cancer: Luteinising Hormone-releasing Hormone Agonists Versus Transdermal Oestradiol. European urology. 2016;69(6):1016-25. doi: 10.1016/j.eururo.2015.11.030.
114. Langley RE, Cafferty FH, Alhasso AA, Rosen SD, Sundaram SK, Freeman SC, et al. Cardiovascular outcomes in patients with locally advanced and metastatic prostate cancer treated with luteinising-hormone-releasing-hormone agonists or transdermal oestrogen: the randomised, phase 2 MRC PATCH trial (PR09). The Lancet Oncology. 2013;14(4):306-16.
115. Langley RE, Gilbert DC, Duong T, Clarke NW, Nankivell M, Rosen SD, et al. Transdermal oestradiol for androgen suppression in prostate cancer: long-term cardiovascular outcomes from the randomised Prostate Adenocarcinoma Transcutaneous Hormone (PATCH) trial programme. The Lancet. 2021;397(10274):581-91. doi: 10.1016/S0140-6736(21)00100-8.

116. Wright-Hughes A, Graham E, Cottrell D, Farrin A. Routine hospital data – is it good enough for trials? An example using England's Hospital Episode Statistics in the SHIFT trial of Family Therapy vs. Treatment as Usual in adolescents following self-harm. Clinical Trials. 2018;15(2):197-206. doi: 10.1177/1740774517751381.
117. NHS England. ECDS 2015. [Available from: https://www.england.nhs.uk/wp-content/uploads/2015/12/ecds-v2-1.pdf.
118. NICOR. About Heart Failure 2020. [Available from: https://www.nicor.org.uk/national-cardiac-audit-programme/about-heart-failure/. Accessed 02/07/2020.
119. NICOR. Myocaridal Ischaemia National Audit Project 2019 Summary Report 2019 [Available from: https://www.nicor.org.uk/wp-content/uploads/2019/09/MINAP-2019-Summary-Report-final.pdf. Accessed 01/05/2020.
120. Macnair A, Love SB, Murray ML, Gilbert DC, Parmar MKB, Denwood T, et al. Accessing routinely collected health data to improve clinical trials: recent experience of access. Trials. 2021;22(1):340. doi: 10.1186/s13063-021-05295-5.
121. NHS Digital. HES Data Dictionary Accident and Emergency. 2020. [Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hospital-episode-statistics-data-dictionary. Accessed 07/12/2020.
122. Harper C, Mafham M, Herrington W, Staplin N, Stevens W, Wallendszus K, et al. Comparison of the Accuracy and Completeness of Records of Serious Vascular Events in Routinely Collected Data vs Clinical Trial–Adjudicated Direct Follow-up Data in the UK: Secondary Analysis of the ASCEND Randomized Clinical Trial. JAMA Network Open. 2021;4(12):e2139748-e. doi: 10.1001/jamanetworkopen.2021.39748
123. Wood A, Denholm R, Hollings S, Cooper J, Ip S, Walker V, et al. Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. BMJ. 2021;373:n826. doi: 10.1136/bmj.n826.
124. Naci H, Davis C, Savović J, Higgins JPT, Sterne JAC, Gyawali B, et al. Design characteristics, risk of bias, and reporting of randomised controlled trials supporting approvals of cancer drugs by European Medicines Agency, 2014-16: cross sectional analysis. BMJ. 2019;366:l5221. doi: 10.1136/bmj.l5221.
125. Elliss-Brookes L, McPhail S, Ives A, Greenslade M, Shelton J, Hiom S, et al. Routes to diagnosis for cancer – determining the patient journey using multiple routine data sets. British Journal of Cancer. 2012;107(8):1220-6. doi: 10.1038/bjc.2012.408.
126. NHS England and Improvement. NHS Operational Planning and Contracting Guidnace. 2020. [Available from: https://www.england.nhs.uk/wp-content/uploads/2020/01/2020-21-NHS-Operational-Planning-Contracting-Guidance.pdf
127. Kwakkenbos L, Juszczak E, Hemkens LG, Sampson M, Fröbert O, Relton C, et al. Protocol for the development of a CONSORT extension for RCTs using cohorts and routinely collected health data. Research Integrity and Peer Review. 2018;3(1):9. doi: 10.1186/s41073-018-0053-3.
128. Fitzpatrick T, Perrier L, Shakik S, Cairncross Z, Tricco AC, Lix L, et al. Assessment of Long-term Follow-up of Randomized Trial Participants by Linkage to Routinely Collected Data: A Scoping Review and Analysis. JAMA Network Open. 2018;1(8):e186019-e. doi: 10.1001/jamanetworkopen.2018.6019.
129. Early Breast Cancer Trialists' Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomised trials. The Lancet. 1998;351(9114):1451-67.

130. Bolla M, Van Tienhoven G, Warde P, Dubois JB, Mirimanoff R-O, Storme G, et al. External irradiation with or without long-term androgen suppression for prostate cancer with high metastatic risk: 10-year results of an EORTC randomised study. The Lancet Oncology. 2010;11(11):1066-73. doi: 10.1016/S1470-2045(10)70223-0.
131. Mason MD, Parulekar WR, Sydes MR, Brundage M, Kirkbride P, Gospodarowicz M, et al. Final Report of the Intergroup Randomized Study of Combined Androgen-Deprivation Therapy Plus Radiotherapy Versus Androgen-Deprivation Therapy Alone in Locally Advanced Prostate Cancer. Journal of clinical oncology. 2015;33(19):2143-50. doi: 10.1200/JCO.2014.57.7510.
132. Statista. Percentage of cumulative clinical trials started worldwide between 2000 and 2018, by cancer type 2020. [Available from: https://www.statista.com/statistics/1092728/cumulative-clinical-trial-starts-share-by-cancer-indication/. Accessed 02/12/2020.
133. Kilburn LS, Banerji J, Bliss JM. The challenges of long-term follow-up data collection in non-commercial, academically-led breast cancer clinical trials: the UK perspective. Trials. 2014;15:379. doi:10.1186/1745-6215-15-379.
134. Hind D, Reeves BC, Bathers S, Bray C, Corkhill A, Hayward C, et al. Comparative costs and activity from a sample of UK clinical trials units. Trials. 2017;18(1):203. doi: 10.1186/s13063-017-1934-3.
135. NHS Improvement. Innovation to implementation: Stratified pathways of care for people living with or beyond cancer – A 'how to guide' 2016. [Available from: https://www.england.nhs.uk/wp-content/uploads/2016/04/stratified-pathways-update.pdf. Accessed 02/02/2020.
136. Llewellyn-Bennett R, Edwards D, Roberts N, Hainsworth AH, Bulbulia R, Bowman L. Post-trial follow-up methodology in large randomised controlled trials: a systematic review. Trials. 2018;19(1):298. doi: 10.1186/s13063-018-2653-0.
137. Fewtrell MS, Kennedy K, Singhal A, Martin RM, Ness A, Hadders-Algra M, et al. How much loss to follow-up is acceptable in long-term randomised trials and prospective studies? Archives of Disease in Childhood. 2008;93(6):458-61. doi: 10.1136/adc.2007.127316.
138. Fergusson D, Aaron SD, Guyatt G, Hébert P. Post-randomisation exclusions: the intention to treat principle and excluding patients from analysis. BMJ. 2002;325(7365):652-4. doi: 10.1136/bmj.325.7365.652.
139. Iyer R, Gentry-Maharaj A, Nordin A, Liston R, Burnell M, Das N, et al. Patient-reporting improves estimates of postoperative complication rates: a prospective cohort study in gynaecological oncology. Br J Cancer. 2013;109(3):623-32. doi: 10.1038/bjc.2013.366.
140. Velikova G, Booth L, Smith AB, Brown PM, Lynch P, Brown JM, et al. Measuring Quality of Life in Routine Oncology Practice Improves Communication and Patient Well-Being: A Randomized Controlled Trial. Journal of clinical oncology. 2004;22(4):714-24. doi: 10.1200/JCO.2004.06.078.
141. Fourkala EO, Gentry-Maharaj A, Burnell M, Ryan A, Manchanda R, Dawnay A, et al. Histological confirmation of breast cancer registration and self-reporting in England and Wales: a cohort study within the UK Collaborative Trial of Ovarian Cancer Screening. Br J Cancer. 2012;106(12):1910-6. doi: 10.1038/bjc.2012.155.
142. Thomas DS, Gentry-Maharaj A, Ryan A, Fourkala E-O, Apostolidou S, Burnell M, et al. Colorectal cancer ascertainment through cancer registries, hospital episode statistics, and self-reporting compared to confirmation by clinician: A cohort study nested within the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). Cancer epidemiology. 2019;58:167-74. doi: 10.1016/j.canep.2018.11.011.

143. Franklin M, Thorn J. Self-reported and routinely collected electronic healthcare resource-use data for trial-based economic evaluations: the current state of play in England and considerations for the future. BMC medical research methodology. 2019;19(1):8. doi: 10.1186/s12874-018-0649-9.
144. Cherry N, McNamee R, Heagerty A, Kitchener H, Hannaford P. Long-term safety of unopposed estrogen used by women surviving myocardial infarction: 14-year follow-up of the ESPRIT randomised controlled trial. BJOG. 2014;121(6):700-5; discussion 705. doi: 10.1111/1471-0528.12598.
145. Cuzick J, Sestak I, Cawthorn S, Hamed H, Holli K, Howell A, et al. Tamoxifen for prevention of breast cancer: extended long-term follow-up of the IBIS-I breast cancer prevention trial. The Lancet Oncology. 2015;16(1):67-75. doi: 10.1016/S1470-2045(14)71171-4.
146. Haynes R, Harden P, Judge P, Blackwell L, Emberson J, Landray MJ, et al. Alemtuzumab-based induction treatment versus basiliximab-based induction treatment in kidney transplantation (the 3C Study): a randomised trial. The Lancet. 2014;384(9955):1684-90. doi: 10.1016/S0140-6736(14)61095-3.
147. 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;319(9):883-95. doi: 10.1001/jama.2018.0154.
148. Murphy PB, Rehal S, Arbane G, Bourke S, Calverley PMA, Crook AM, et al. Effect of Home Noninvasive Ventilation With Oxygen Therapy vs Oxygen Therapy Alone on Hospital Readmission or Death After an Acute COPD Exacerbation: A Randomized Clinical Trial. JAMA. 2017;317(21):2177-86. doi: 10.1001/jama.2017.4451.
149. Herrington W, Wallendszus K, Bowman L, Landray M, Armitage J. Can vascular mortality be reliably ascertained from the underlying cause of death recorded on a medical death certificate? Evidence from 2800 adjudicated heart protection study (HPS) deaths. Trials. 2015;16(2):P61. doi: 10.1186/1745-6215-16-S2-P61.
150. Turner EL, Metcalfe C, Donovan JL, Noble S, Sterne JA, Lane JA, et al. Contemporary accuracy of death certificates for coding prostate cancer as a cause of death: Is reliance on death certification good enough? A comparison with blinded review by an independent cause of death evaluation committee. Br J Cancer. 2016;115(1):90-4. doi: 10.1038/bjc.2016.162.
151. Kilburn LS, Aresu M, Banerji J, Barrett-Lee P, Ellis P, Bliss JM. Can routine data be used to support cancer clinical trials? A historical baseline on which to build: retrospective linkage of data from the TACT (CRUK 01/001) breast cancer trial and the National Cancer Data Repository. Trials. 2017;18(1):561. doi: 10.1186/s13063-017-2308-6.
152. Love SB, Kilanowski A, Yorke-Edwards V, Old O, Barr H, Stokes C, et al. Use of routinely collected health data in randomised clinical trials: comparison of trial-specific death data in the BOSS trial with NHS Digital data. Trials. 2021;22(1):654. doi: 10.1186/s13063-021-05613-x.
153. Bhattacharya IS, Morden JP, Griffin C, Snowdon C, Brannan R, Bliss JM, et al. The Application and Feasibility of Using Routine Data Sources for Long-term Cancer Clinical Trial Follow-up. Clin Oncol. 2017;29(12):796-8. doi: 10.1016/j.clon.2017.10.007.
154. Kalsi JK, Ryan A, Gentry-Maharaj A, Margolin-Crump D, Singh N, Burnell M, et al. Completeness and accuracy of national cancer and death registration for

outcome ascertainment in trials—an ovarian cancer exemplar. Trials. 2021;22(1):88. doi: 10.1186/s13063-020-04968-x.
155. National Cancer Registration and Analysis Service. 2014-2017 Recurrence Reporting by Hospital Trust. 2017. [Available from : http://ncin.org.uk/cancer_type_and_topic_specific_work/topic_specific_work/recurrence. Accessed 02/12/2020.
156. Cameron D, Morden JP, Canney P, Velikova G, Coleman R, Bartlett J, et al. Accelerated versus standard epirubicin followed by cyclophosphamide, methotrexate, and fluorouracil or capecitabine as adjuvant therapy for breast cancer in the randomised UK TACT2 trial (CRUK/05/19): a multicentre, phase 3, open-label, randomised, controlled trial. The Lancet Oncol. 2017;18(7):929-45. doi: 10.1016/S1470-2045(17)30404-7.
157. Coyle C, Cafferty FH, Rowley S, MacKenzie M, Berkman L, Gupta S, et al. ADD-ASPIRIN: A phase III, double-blind, placebo controlled, randomised trial assessing the effects of aspirin on disease recurrence and survival after primary therapy in common non-metastatic solid tumours. Contemp Clin Trials. 2016;51:56-64. doi: 10.1016/j.cct.2016.10.004.
158. Rothwell PM, Fowkes FG, Belch JF, Ogawa H, Warlow CP, Meade TW. Effect of daily aspirin on long-term risk of death due to cancer: analysis of individual patient data from randomised trials. The Lancet. 2011;377(9759):31-41. doi: 10.1016/S0140-6736(10)62110-1.
159. Rothwell PM, Wilson M, Price JF, Belch JF, Meade TW, Mehta Z. Effect of daily aspirin on risk of cancer metastasis: a study of incident cancers during randomised controlled trials. The Lancet. 2012;379(9826):1591-601. doi: 10.1016/S0140-6736(12)60209-8.
160. SWAT 125: Comparison of trial-collected and routinely-collected death data 2020. [Available from: https://www.qub.ac.uk/sites/TheNorthernIrelandNetworkforTrialsMethodologyResearch/FileStore/Filetoupload,976743,en.pdf.
161. Hippisley-Cox J, Coupland C. Predicting risk of upper gastrointestinal bleed and intracranial bleed with anticoagulants: cohort study to derive and validate the QBleed scores. BMJ (Clinical research ed). 2014;349:g4606. doi: 10.1136/bmj.g4606.
162. Public Health England. ODR approval guidelines: cost recovery. 2020. [Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/886163/ODR_Approval_guidelines_cost_recovery_2021.pdf
163. Xu Y, Kong S, Cheung WY, Bouchard-Fortier A, Dort JC, Quan H, et al. Development and validation of case-finding algorithms for recurrence of breast cancer using routinely collected administrative data. BMC Cancer. 2019;19(1):210. doi: 10.1186/s12885-019-5432-8.
164. Colov EP, Fransgaard T, Klein M, Gögenur I. Validation of a register-based algorithm for recurrence in rectal cancer. Dan Med J. 2018;65(10).
165. Hassett MJ, Uno H, Cronin AM, Carroll NM, Hornbrook MC, Ritzwoller D. Detecting Lung and Colorectal Cancer Recurrence Using Structured Clinical/Administrative Data to Enable Outcomes Research and Population Health Management. Med Care. 2017;55(12):e88-e98. doi: 10.1097/MLR.0000000000000404.
166. Ritzwoller DP, Hassett MJ, Uno H, Cronin AM, Carroll NM, Hornbrook MC, et al. Development, Validation, and Dissemination of a Breast Cancer Recurrence

Detection and Timing Informatics Algorithm. J Natl Cancer Inst. 2018;110(3):273-81. doi: 10.1093/jnci/djx200.
167. National Cancer Registration and Analysis Service. Rapid Cancer Registry Database 2020. [Available from: http://www.ncin.org.uk/collecting_and_using_data/rcrd. Accessed 12/01/2021
168. Public Health Scotland. Scottish Cancer Registry 2020. [Available from: https://www.isdscotland.org/Health-Topics/Cancer/Scottish-Cancer-Registry/How-data-are-collected/. Accessed 11/01/2021.
169. Woodcock A, Vestbo J, Bakerly ND, New J, Gibson JM, McCorkindale S, et al. Effectiveness of fluticasone furoate plus vilanterol on asthma control in clinical practice: an open-label, parallel group, randomised controlled trial. The Lancet. 2017;390(10109):2247-55. doi: 10.1016/S0140-6736(17)32397-8.
170. New JP, Bakerly ND, Leather D, Woodcock A. Obtaining real-world evidence: the Salford Lung Study. Thorax. 2014;69(12):1152-4. doi: 10.1136/thoraxjnl-2014-205259.
171. Plescia OJ, Smith AH, Grinwich K. Subversion of immune system by tumor cells and role of prostaglandins. Proc Natl Acad Sci USA. 1975;72(5):1848-51.
172. Schrör K. Pharmacology and cellular/molecular mechanisms of action of aspirin and Non-aspirin NSAIDs in colorectal cancer. Best Practice & Research Clinical Gastroenterology. 2011;25(4):473-84. doi: 10.1016/j.bpg.2011.10.016.
173. Thun MJ, Jacobs EJ, Patrono C. The role of aspirin in cancer prevention. Nature Reviews Clinical Oncology. 2012;9(5):259-67. doi: 10.1038/nrclinonc.2011.199.
174. Davì G, Patrono C. Platelet Activation and Atherothrombosis. New England Journal of Medicine. 2007;357(24):2482-94. doi: 10.1056/NEJMra071014.
175. Guillem-Llobat P, Dovizio M, Bruno A, Ricciotti E, Cufino V, Sacco A, et al. Aspirin prevents colorectal cancer metastasis in mice by splitting the crosstalk between platelets and tumor cells. Oncotarget. 2016;7(22):32462-77. doi: 10.18632/oncotarget.8655.
176. Elwood PC, Gallagher AM, Duthie GG, Mur LA, Morgan G. Aspirin, salicylates, and cancer. The Lancet. 2009;373(9671):1301-9. doi: 10.1016/S0140-6736(09)60243-9.
177. Chan AT, Ogino S, Fuchs CS. Aspirin and the Risk of Colorectal Cancer in Relation to the Expression of COX-2. New England Journal of Medicine. 2007;356(21):2131-42. doi: 10.1056/NEJMoa067208.
178. Bosetti C, Rosato V, Gallus S, Cuzick J, La Vecchia C. Aspirin and cancer risk: a quantitative review to 2011. Annals of Oncology. 2012;23(6):1403-15. doi: 10.1093/annonc/mds113.
179. Bosetti C, Santucci C, Gallus S, Martinetti M, La Vecchia C. Aspirin and the risk of colorectal and other digestive tract cancers: updated meta-analysis through 2019. Annals of Oncology. 2020;31(5):558-68. doi: 10.1016/j.annonc.2020.02.012.
180. Cuzick J, Thorat MA, Bosetti C, Brown PH, Burn J, Cook NR, et al. Estimates of benefits and harms of prophylactic use of aspirin in the general population. Annals of Oncology. 2015;26(1):47-57. doi: 10.1093/annonc/mdu225.
181. Qiao Y, Yang T, Gan Y, Li W, Wang C, Gong Y, et al. Associations between aspirin use and the risk of cancers: a meta-analysis of observational studies. BMC Cancer. 2018;18(1):288. doi: 10.1186/s12885-018-4156-5.
182. Rothwell PM, Price JF, Fowkes FG, Zanchetti A, Roncaglioni MC, Tognoni G, et al. Short-term effects of daily aspirin on cancer incidence, mortality, and non-vascular death: analysis of the time course of risks and benefits in 51 randomised

controlled trials. The Lancet. 2012;379(9826):1602-12. doi: 10.1016/S0140-6736(11)61720-0.
183. Gaziano JM, Brotons C, Coppolecchia R, Cricelli C, Darius H, Gorelick PB, et al. Use of aspirin to reduce risk of initial vascular events in patients at moderate risk of cardiovascular disease (ARRIVE): a randomised, double-blind, placebo-controlled trial. The Lancet. 2018;392(10152):1036-46. doi: 10.1016/S0140-6736(18)31924-X.
184. McNeil JJ, Nelson MR, Woods RL, Lockery JE, Wolfe R, Reid CM, et al. Effect of Aspirin on All-Cause Mortality in the Healthy Elderly. New England Journal of Medicine. 2018;379(16):1519-28. doi: 10.1056/NEJMoa1803955.
185. The ASCEND Study Collaborative Group. Effects of Aspirin for Primary Prevention in Persons with Diabetes Mellitus. New England Journal of Medicine. 2018;379(16):1529-39. doi: 10.1056/NEJMoa1804988.
186. Wu Q, Yao X, Chen H, Liu Z, Li T, Fan X, et al. Long-term aspirin use for primary cancer prevention: An updated systematic review and subgroup meta-analysis of 29 randomized clinical trials. J Cancer. 2020;11(21):6460-73. doi: 10.7150/jca.49001.
187. Cook NR, Lee IM, Zhang SM, Moorthy MV, Buring JE. Alternate-day, low-dose aspirin and cancer risk: long-term observational follow-up of a randomized trial. Ann Intern Med. 2013;159(2):77-85. doi: 10.7326/0003-4819-159-2-201307160-00002.
188. Flossmann E, Rothwell PM. Effect of aspirin on long-term risk of colorectal cancer: consistent evidence from randomised and observational studies. The Lancet. 2007;369(9573):1603-13. doi: 10.1016/S0140-6736(07)60747-8.
189. Rothwell PM, Wilson M, Elwin CE, Norrving B, Algra A, Warlow CP, et al. Long-term effect of aspirin on colorectal cancer incidence and mortality: 20-year follow-up of five randomised trials. The Lancet. 2010;376(9754):1741-50. doi: 10.1016/S0140-6736(10)61543-7.
190. Burn J, Gerdes AM, Macrae F, Mecklin JP, Moeslein G, Olschwang S, et al. Long-term effect of aspirin on cancer risk in carriers of hereditary colorectal cancer: an analysis from the CAPP2 randomised controlled trial. The Lancet. 2011;378(9809):2081-7. doi: 10.1016/S0140-6736(11)61049-0.
191. Algra AM, Rothwell PM. Effects of regular aspirin on long-term cancer incidence and metastasis: a systematic comparison of evidence from observational studies versus randomised trials. The Lancet Oncology. 2012;13(5):518-27. doi: 10.1016/S1470-2045(12)70112-2.
192. Chubak J, Whitlock EP, Williams SB, Kamineni A, Burda BU, Buist DS, et al. Aspirin for the Prevention of Cancer Incidence and Mortality: Systematic Evidence Reviews for the U.S. Preventive Services Task Force. Ann Intern Med. 2016;164(12):814-25. doi: 10.7326/M15-2117.
193. Mahase E. US taskforce advises against low dose aspirin for primary prevention of cardiovascular disease. BMJ. 2021;375:n2521. doi:10.1136/bmj.n2521.
194. 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;397(10290):2182-93. doi: 10.1016/S0140-6736(21)00731-5.
195. Simon TG, Duberg A-S, Aleman S, Chung RT, Chan AT, Ludvigsson JF. Association of Aspirin with Hepatocellular Carcinoma and Liver-Related Mortality. New England Journal of Medicine. 2020;382(11):1018-28. doi: 10.1056/NEJMoa1912035.

196. Chan AT, Giovannucci EL, Meyerhardt JA, Schernhammer ES, Wu K, Fuchs CS. Aspirin Dose and Duration of Use and Risk of Colorectal Cancer in Men. Gastroenterology. 2008;134(1):21-8. doi: 10.1053/j.gastro.2007.09.035.
197. Simon TG, Ma Y, Ludvigsson JF, Chong DQ, Giovannucci EL, Fuchs CS, et al. Association Between Aspirin Use and Risk of Hepatocellular Carcinoma. JAMA oncology. 2018;4(12):1683-90. doi: 10.1001/jamaoncol.2018.4154.
198. Loomans-Kropp HA, Pinsky P, Cao Y, Chan AT, Umar A. Association of Aspirin Use With Mortality Risk Among Older Adult Participants in the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial. JAMA Netw Open. 2019;2(12):e1916729. doi: 10.1001/jamanetworkopen.2019.16729.
199. Downer MK, Allard CB, Preston MA, Wilson KM, Kenfield SA, Chan JM, et al. Aspirin Use and Lethal Prostate Cancer in the Health Professionals Follow-up Study. European Urology Oncology. 2019;2(2):126-34. doi: 10.1016/j.euo.2018.07.002.
200. Hurwitz LM, Michels KA, Cook MB, Pfeiffer RM, Trabert B. Associations between daily aspirin use and cancer risk across strata of major cancer risk factors in two large U.S. cohorts. Cancer Causes & Control. 2021;32(1):57-65. doi: 10.1007/s10552-020-01357-2.
201. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393. doi: 10.1136/bmj.b2393.
202. Gray RT, Coleman HG, Hughes C, Murray LJ, Cardwell CR. Low-dose aspirin use and survival in colorectal cancer: results from a population-based cohort study. BMC Cancer. 2018;18(1):228. doi: 10.1186/s12885-018-4142-y.
203. McNeil JJ, Woods RL, Nelson MR, Reid CM, Kirpach B, Wolfe R, et al. Effect of Aspirin on Disability-free Survival in the Healthy Elderly. New England Journal of Medicine. 2018;379(16):1499-508. doi: 10.1056/NEJMoa1800722.
204. Burn J, Sheth H, Elliott F, Reed L, Macrae F, Mecklin JP, et al. Cancer prevention with aspirin in hereditary colorectal cancer (Lynch syndrome), 10-year follow-up and registry-based 20-year data in the CAPP2 study: a double-blind, randomised, placebo-controlled trial. The Lancet. 2020;395(10240):1855-63. doi: 10.1016/S0140-6736(20)30366-4.
205. National Institute for Health and Care Excellance. Lynch syndrome: should I take aspirin to reduce my chance of getting bowel cancer? 2020. [Available from: https://www.nice.org.uk/guidance/ng151/resources/patient-decision-aid-pdf-8834927870. Accessed 12/12/21
206. NHS Digital. General Practice for data and research 2021. [Available from: https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/general-practice-data-for-planning-and-research. Accessed 07/06/2021.
207. Powell GA, Bonnett LJ, Smith CT, Hughes DA, Williamson PR, Marson AG. Using routinely recorded data in a UK RCT: a comparison to standard prospective data collection methods. Trials. 2021;22(1):429. doi: 10.1186/s13063-021-05294-6
208. Health Data Research UK. Health Data Research Innovation Gateway 2020. [Available from: https://www.healthdatagateway.org/. Accessed 06/08/2020.
209. Sydes MR, Barbachano Y, Bowman L, Denwood T, Farmer A, Garfield-Birkbeck S, et al. Realising the full potential of data-enabled trials in the UK: a call for action. BMJ Open. 2021;11(6):e043906. doi: 10.1136/ bmjopen-2020-043906.
210. The RECOVERY collaborative group. Dexamethasone in Hospitalized Patients with Covid-19. New England Journal of Medicine. 2020;384(8):693-704. doi: 10.1056/NEJMoa2021436.

211. Public Health England. Simulacrum 2021. [Available from: https://www.cancerdata.nhs.uk/simulacrum. Accessed 22/12/21.

Appendix A: PATCH and Add-Aspirin transparency statements

PATCH trial Transparency statement

How we use your data

The co-ordination of the study has not changed since it began however there have been new data protection regulations introduced across the UK and Europe and we need to update all of our PATCH participants on how we collect and protect the information provided for the study.

The MRC CTU at UCL will continue to use information from you and / or your medical records in order to undertake this study and will act as the joint data controller for this study with the Sponsor. This means that we are responsible for looking after your information and using it properly.

Your rights to access, change or move your information are limited as we need to manage your information in specific ways in order for the research to be reliable and accurate. If you withdraw from the study, we will keep the information about you that we have already obtained. To safeguard your rights, we will use the minimum personally-identifiable information possible.

You can find out more about how we use your information at www.ctu.mrc.ac.uk/general/privacy-policy.

How your data will be stored and collected

Your hospital will collect information from you and from your medical records for this research study in accordance with our instructions. They will use this information as needed for your care.

The MRC CTU at UCL will collect information about you for this research study from your hospital. This information will include health information, which is regarded as a special category of information. We will use this information to conduct our research. This information is supplied via paper forms designed to collect the study information.

Certain individuals from the MRC CTU at UCL, the Sponsor and regulatory organisations may look at your medical and research records to check the accuracy of the research study. UCL will only receive information without any identifying information. The people who analyse the information will not be able to identify you and will not be able to find out your name, NHS number or contact details.

The MRC CTU at UCL will keep information about you for a minimum of 25 years after the study has finished.

Trial Participant Information Linkage

The information we receive from study staff at your hospital provides the PATCH researchers with information about your progress. However we need accurate long term information to know if the treatments being tested are improving life expectancy.

Through our research we would like to improve reliability of the collection of long term data of a study's result by investigating linkage of the information we collect about you with electronic health records held by a variety of national registries and bodies such as the Office of National Statistics, NHS Digital, Public Health England and National Clinical Audit programs.

We will securely transfer the directly identifiable data such as your name, NHS number or postcode for this purpose only. This is information that you provided when you first joined the study to enable us to do this. We will continue to store this personal data separately from the clinical data. All information is stored securely at the MRC CTU at UCL and the data controller is UCL. Only the mentioned parties will access the identifiable information on the participants (ONS, NHS Digital, national Clinical Audit programs and UCL). Any published results from the trial will not lead to participants being directly identified.

If at any point you do not want us to collect information about your health from national sources of health information then please talk to your study doctor or nurse who will then inform the PACTH study team of your decision. Contact details for members of your study team will be listed within your Patient Information Sheet (PIS). This decision will not affect the care you receive in any way.

How your data will be used in future & other research

When you agree to take part in a research study, the information about your health and care may be provided to researchers running other research studies in this organisation and in other organisations. These organisations may be universities, NHS organisations or companies involved in health and care research in this country or abroad. Your information will only be used by organisations and researchers to conduct research in accordance with the relevant legislation, ethics and research policy requirements.

This information will not identify you and will not be combined with other information in a way that could identify you. The information will only be used for the purpose of health and care research and cannot be used to contact you or to affect your care. It will not be used to make decisions about future services available to you, such as insurance.

If you have any questions or concerns about the Study please contact your PATCH site team.

Add-Aspirin Transparency statement

The Add Aspirin study is aiming to find out whether taking aspirin daily for 5 years after treatment for an early stage cancer (cancer that has not spread widely), stops or delays the cancer coming back. University College London (UCL), based in the United Kingdom, is the sponsor for this study in the United Kingdom and Republic of Ireland (Tata Memorial Centre is the sponsor for the trial in India). University College London, through the Medical Research Council (MRC) Clinical Trial Unit at UCL, will be using information from you and your medical records in order to undertake this study and will act as data controller for this study. UCL will be responsible for looking after your information and using it properly. UCL will keep identifiable information about you for 25 years after the study has finished.

Your rights to access, change, or move your information, are limited; as we need to manage your information in specific ways in order for the research to be reliable and accurate. If you withdraw from the study, we will keep the information about you that we have already obtained. To safeguard your rights, we will use the minimum personally – identifiable information possible.

You can find out more about how we use your information at www.ctu.mrc.ac.uk/general/privacy-policy

Your site will collect information from you and your medical records for this research study in accordance with our instructions.

For UK Participants Only:

Your hospital will use your name, NHS number and contact details to contact you about the research study, and make sure that relevant information about the study is recorded for your care, and to oversee the quality of the study. Individuals from UCL and regulatory organizations may look at your medical and research records to check the accuracy of the research study. Your hospital will pass your name, postcode and NHS number to UCL along with the information collected from you and your medical records. The people who analyse the information will not be able to identify you and will not be able to find out your name, NHS number or contact details.

Your hospital will keep identifiable information about you from this study for at least 25 years after the study has finished.

UCL will collect information about you, for research, from your hospital site, NHS

Digital, Public Health England (PHE) and the National Cancer Registration and Analysis Service (NCRAS). This information will include your name, postcode and NHS number and health information. This health information is regarded as a special category of information as defined by the General Data Protection Regulation (GDPR). The legal basis for collection of sensitive and personal data is it will be used for research purposes in a task in the public interest. We will use this information collected from PHE to track your long term health status and you can find out more information about PHE at http://www.ncin.org.uk/collecting_and_using_data/.

Where information could identify you, the information will be held securely with strict arrangements about who can access the information.

Appendix B: Accessing routinely collected health data to improve clinical trials: recent experience of access

Macnair *et al. Trials* (2021) 22:340
https://doi.org/10.1186/s13063-021-05295-5

Trials

RESEARCH Open Access

Accessing routinely collected health data to improve clinical trials: recent experience of access

Archie Macnair[1,2*], Sharon B. Love[1,2], Macey L. Murray[1,2], Duncan C. Gilbert[1], Mahesh K. B. Parmar[1], Tom Denwood[3], James Carpenter[1,2,4], Matthew R. Sydes[1,2], Ruth E. Langley[1†] and Fay H. Cafferty[1†]

Abstract

Background: Routinely collected electronic health records (EHRs) have the potential to enhance randomised controlled trials (RCTs) by facilitating recruitment and follow-up. Despite this, current EHR use is minimal in UK RCTs, in part due to ongoing concerns about the utility (reliability, completeness, accuracy) and accessibility of the data. The aim of this manuscript is to document the process, timelines and challenges of the application process to help improve the service both for the applicants and data holders.

Methods: This is a qualitative paper providing a descriptive narrative from one UK clinical trials unit (MRC CTU at UCL) on the experience of two trial teams' application process to access data from three large English national datasets: National Cancer Registration and Analysis Service (NCRAS), National Institute for Cardiovascular Outcomes Research (NICOR) and NHS Digital to establish themes for discussion. The underpinning reason for applying for the data was to compare EHRs with data collected through case report forms in two RCTs, Add-Aspirin (ISRCTN 74358648) and PATCH (ISRCTN 70406718).

Results: The Add-Aspirin trial, which had a pre-planned embedded sub-study to assess EHR, received data from NCRAS 13 months after the first application. In the PATCH trial, the decision to request data was made whilst the trial was recruiting. The study received data after 8 months from NICOR and 15 months for NHS Digital following final application submission. This concluded in May 2020. Prior to application submission, significant time and effort was needed particularly in relation to the PATCH trial where negotiations over consent and data linkage took many years.

Conclusions: Our experience demonstrates that data access can be a prolonged and complex process. This is compounded if multiple data sources are required for the same project. This needs to be factored in when planning to use EHR within RCTs and is best considered prior to conception of the trial. Data holders and researchers are endeavouring to simplify and streamline the application process so that the potential of EHR can be realised for clinical trials.

Keywords: Routinely collected data, Electronic health records, Data accessibility, Clinical trials

* Correspondence: a.macnair@ucl.ac.uk
†Ruth E. Langley and Fay H. Cafferty contributed equally to this work.
[1]MRC Clinical Trials Unit at UCL, UCL, London WC1V 6LJ, UK
[2]Health Data Research UK, London, UK
Full list of author information is available at the end of the article

Background

Routinely collected electronic health records (EHRs) have been identified as an important innovation in the conduct of randomised clinical trials (RCTs) [1]. EHRs could improve the efficiency and cost of trials by possibly enhancing recruitment, more complete data sets and minimal loss to follow-up [2, 3]. For example, the TASTE trial (ISRCTN16716833), using the Swedish angiography and angioplasty registry, is one of several trials demonstrating the utility of registry-held EHRs to recruit and follow up participants. This study was able to recruit 82% of eligible patients from the registry and obtained complete follow-up data in a trial of 7244 patients [4]. They also demonstrated meaningfully lower costs for managing the study with a cost per participant in the order of ~$50 compared to costs for a conventional RCT which may be in excess of $1000 per participant [4, 5].

EHRs are often collected by centralised registries and audits (national or regional) for purposes other than clinical research to gather detailed information on specific diseases, treatments or populations. However, there are concerns, depending on the source, that data collected in this way may not be of appropriate detail or quality for use in clinical trials [6]. Access to EHRs by researchers usually requires a formal application to the data holder where specific criteria must be evidenced including compliance with information governance (IG) regulations and a clear purpose and legal basis for the data access.

One potential concern for clinical trialists is that the application process will be complex and lengthy and that the data will not be obtained in a timely manner [7]. There have been reports that RCTs were unable to publish trial results due to data access [8]. One example is the EPOCH trial (ISRCTN80682973), where the research team were unable to procure mortality from Welsh data following hospital admissions. As a result, the researchers had to change their planned primary analysis to make sure their publication was not delayed significantly [9].

The aim of this article is to share and reflect upon our experience at the MRC Clinical Trials Unit at UCL (hereafter 'MRC CTU') in applying to three national holders of EHR datasets in the UK for data relating to two ongoing RCTs. The intention is to highlight some of the hurdles in obtaining data and discuss possible solutions. The overarching aim is to assist future applicants and help data providers, who are commonly trying to improve their processes and address these issues in a way that is mutually beneficial.

Methods

This is a qualitative study based on recent experience of the teams at an accredited clinical trials unit (MRC CTU) in applying for and accessing routine datasets in England (for two separate trials). The data access applications are linked by one main applicant as part of their clinical methodology research and use a descriptive narrative from documented exchanges between the data holder and applicant to establish themes for discussion. These were chosen as they cover recent access to some of the main datasets likely to be used by clinical trialists with a range of common clinical outcomes. The MRC CTU sought English EHR data for the Add-Aspirin (ISRCTN 74358648) and PATCH (ISRCTN 70406718) trials.

Add-Aspirin aims to assess whether daily aspirin use after treatment for an early-stage cancer can prevent recurrence and improve survival [10]. It will recruit 11,000 participants in the UK, Republic of Ireland and India; recruitment began in October 2015 and is ongoing. The Add-Aspirin protocol includes a methodological substudy designed to assess the feasibility of applying for and using EHRs from the National Cancer Registration and Analysis Service (NCRAS) [11] to assist in the long-term follow-up of participants after completion of trial treatment.

PATCH is a randomised trial of approximately 2500 participants with prostate cancer in the UK. It is assessing the efficacy and safety of a novel therapy transdermal oestradiol patches against standard hormone therapy [12]. Transdermal patches may have a better side-effect profile compared with standard treatment but there was a prior concern about increased cardiovascular toxicity based on trials of oral oestrogens in the 1970s. PATCH therefore had enhanced monitoring of cardiovascular outcomes, gathering all available information about each event with an additional clinical review [12]. After the trial started, a methodology substudy was initiated to compare serious adverse cardiovascular events reported by research staff at participating sites through trial-specific data collection forms with those routinely collected from, and reported in, audits held by the National Institute for Cardiovascular Outcomes Research (NICOR) and Hospital Episodes Statistics (HES) held by NHS Digital. Concordance between the three datasets would support the premise that routinely collected data could supplement or replace long-term cardiotoxicity data in this trial and other future RCTs.

The routine data to be accessed for these two projects are held and collated by three different organisations with their own individual processes to allow data access. Although the organisations are all within the auspices of the English National Health Service, each has evolved in recent years. This, along with revisions to the legal framework for IG, means that the process of data access has also evolved.

National Cancer Registration and Analysis Service (NCRAS)

In 2016, NCRAS was formed from the merger of the National Cancer Intelligence Network (NCIN) and National Disease Registration (NDR) within Public Health England [13]. In England, NCRAS manages the collection of data relating to cancer. The aim is to monitor cancer incidence, improve care and clinical outcomes, aid research and support genetic counselling [11]. NCRAS hold several different datasets covering cancer registration and cancer treatments (systemic therapy and radiotherapy). They can also link these datasets to others held by NHS Digital or the Office for National Statistics (ONS), such as mortality data and HES, via NHS number or other personal identifiers.

To gain access to this data for research, an application must be submitted to the Office for Data Release (ODR) [14]. The ODR application process is outlined in Fig. 1 [14].

NHS Digital

NHS Digital has been the custodian of HES since 2016. Prior to this, it operated under the Health and Social Care Information Centre (HSC-IC) from 2005 [15]. NHS Digital collects, processes and provides access to many EHR datasets and is continually seeking to supplement this data with other datasets from various care settings. HES is primarily a resource for reimbursement of hospital activity and holds patient-level information on more than 500 variables ranging from diagnosis, procedures, admission dates, demographics of the patients and healthcare provider [16]. NHS Digital has a large number of organisations requesting access to their data with most coming from local authorities and Clinical Commissioning Groups [8]; access is provided by application to the Data Access Request Service (DARS) [17]. The Independent Group Advising on the Release of Data (IGARD) gives an independent final review that aims to improve transparency, accountability, quality and consistency of the application process. IGARD currently meets weekly to make sure that applications are reviewed in a timely fashion. The application process continues to change with attempts to improve its service; the current process is outlined in Fig. 2 [17].

National Institute for Cardiovascular Outcomes Research (NICOR)

NICOR collects routine EHR data and produces analyses to enable hospitals and healthcare improvement bodies to monitor and improve the care and outcomes of patients with cardiovascular disease. It manages six national clinical audits and a number of new health technology registries [18]. NICOR is regulated and contracted by the Health Quality Improvement Partnership (HQIP). NICOR was originally hosted by UCL but moved to Barts Health NHS Trust in 2017. The two audits that were identified as potentially relevant to the PATCH trial were the National Heart Failure Audit (NHFA) and the Myocardial Ischaemia National Audit Project (MINAP). The application process to obtain data from NICOR is shown in Fig. 3 [18]. Historically, far fewer researchers have used this source compared to NHS Digital and NCRAS [8].

Findings

Add-Aspirin

The Add-Aspirin trial was conceived with the recognition that participants will require follow-up for at least 10 years [10]. This length of follow-up is required to assess the overall risk: benefit of regular aspirin use on the

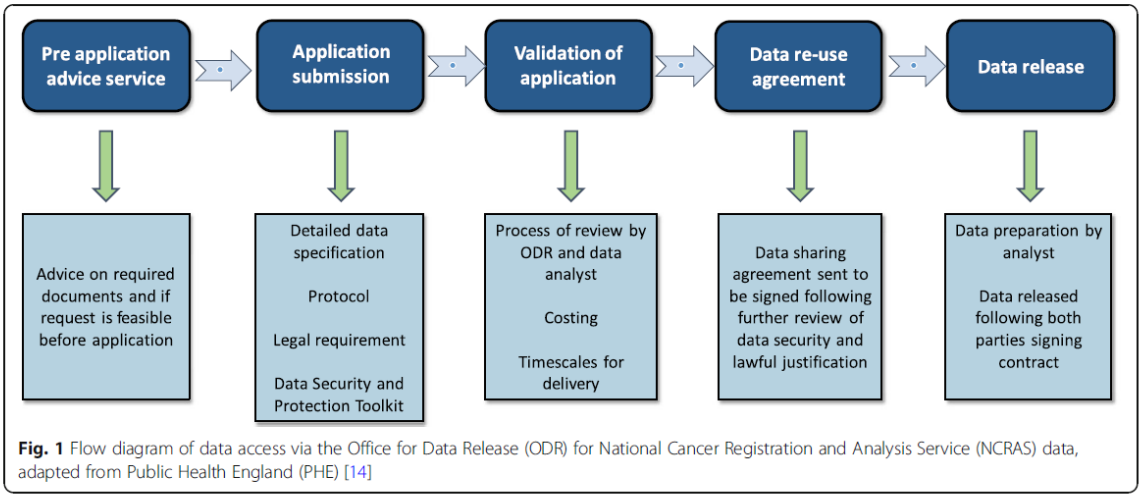


Fig. 1 Flow diagram of data access via the Office for Data Release (ODR) for National Cancer Registration and Analysis Service (NCRAS) data, adapted from Public Health England (PHE) [14]

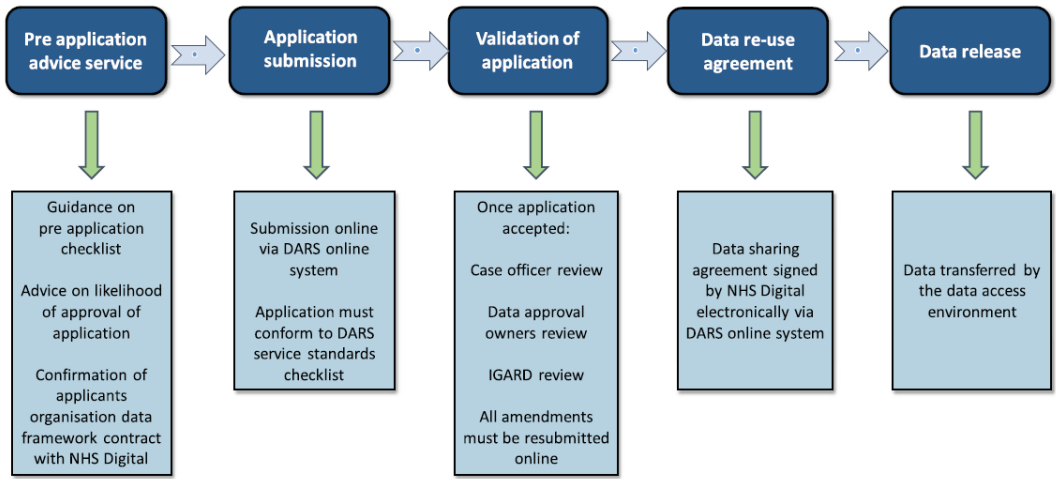


Fig. 2 Flow diagram of the Data Access Request Service (DARS) for NHS Digital data, adapted from NHS Digital [17]. IGARD, Independent Group Advising on the Release of Data

trial participants' health. From the design stage of the trial, like for many trials [19], there was an intention to access data using routinely collected EHRs. When the trial was initially conceived in 2012, the Add-Aspirin trial team met with individuals from NCIN, the predecessor of NCRAS, to assess the feasibility of accessing data and also to ensure that an appropriate budget for this activity was incorporated into funding applications (Fig. 4). The protocol, patient information sheets and consent forms were designed to reflect the potential use of routinely collected healthcare data.

In 2017, after 2 years of recruitment and follow-up, there was a conversation with ODR to confirm the cost and current application process. In 2018, there was sufficient data to initiate the pre-defined methodology substudy. A pre-application meeting with an ODR senior

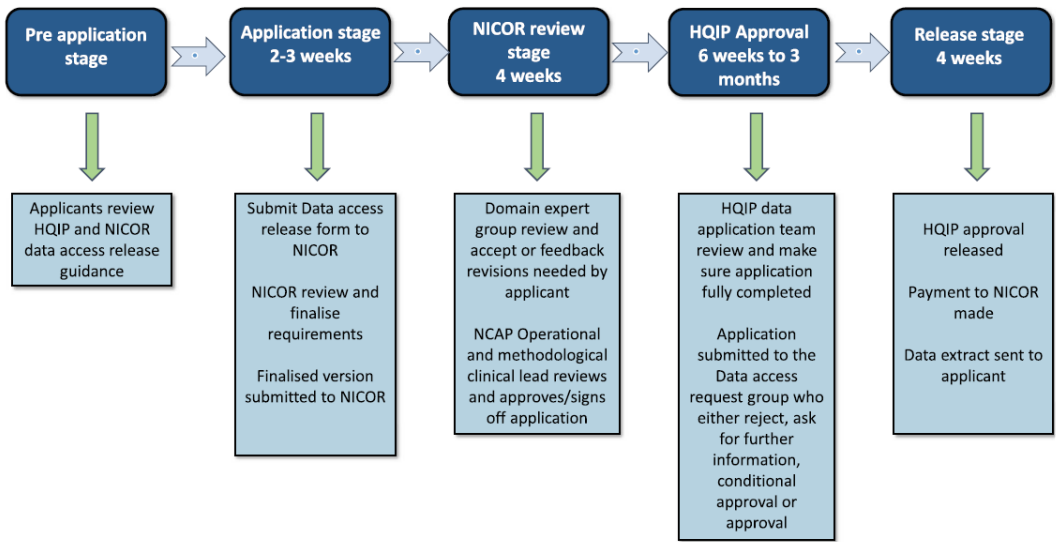


Fig. 3 Data access request for access to National Institute for Cardiovascular Outcomes Research (NICOR) data adapted from [18]. HQIP, Health Quality Improvement Partnership; NCAP, National Cardiac Audit Programme

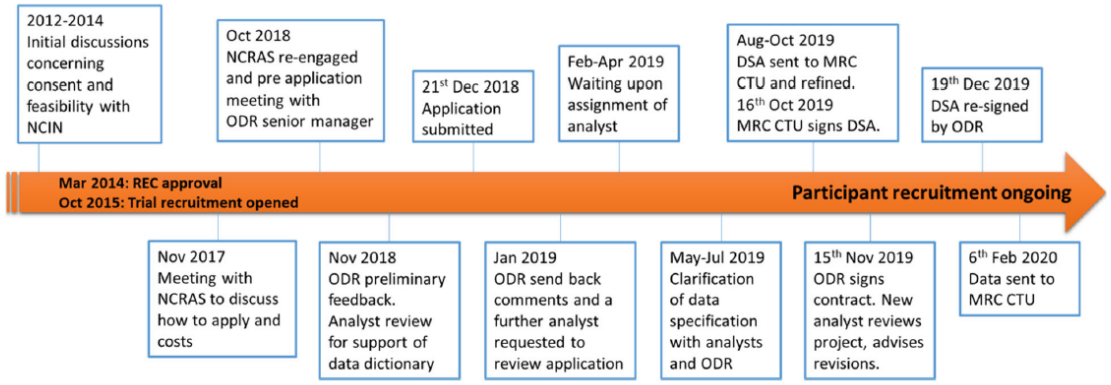


Fig. 4 Flow diagram of the Add-Aspirin National Cancer Registration and Analysis Service (NCRAS) application. (Please note that timeline is not proportional) REC approval for Add-Aspirin March 2014. Recruitment opened in October 2015 and is ongoing. CTU, clinical trials unit; DSA, data sharing agreement; NCIN, National Cancer Intelligence Network; ODR, Office for Data Release; REC, Research Ethics Committee

manager established the documentation that was needed going forward.

Following the implementation of the General Data Protection Regulation (GDPR) in the UK (2018), transparency of how exactly participant data would be used became a legal requirement. The previously agreed consent forms and patient information sheets did not meet the 2018 requirements of GDPR. The solution was for a privacy notice to be drafted and made publicly accessible on the trial's website. The trial's IG documentation also needed updating to ensure information security assurances (via the Data Security and Protection Toolkit) were in place within UCL.

Following submission of the data application (December 2018), ODR sent back revisions (January 2019) and confirmed the transparency statement (February 2019). For the application to proceed, an analyst needed to be allocated to check the defined data requirements. In April 2019, NCRAS unfortunately unassigned the analyst allocated to Add-Aspirin onto work on a project considered more critical. There was a meeting in May 2019, once further analytical support had been deployed, to discuss the data field requests. The new analysts suggested that a number of data fields should be expanded to give the best chance of capturing cancer recurrence as this is not, at present, collected sufficiently well within any single EHR dataset. They acknowledged at that time that algorithms were needed to identify data patterns indicative of tumour recurrence. ODR wanted to ensure that no unnecessary data from HES was provided for each participant. The MRC CTU therefore provided surgical/procedure codes (using Office of Population Censuses and Surveys (OPCS) definitions) and diagnosis codes (ICD-10 codes) to NCRAS to focus and limit the data extraction. In June 2019, it was agreed with ODR and NCRAS that, as this was a methodological project reviewing ways to gather trial outcomes in registry data, all HES data for these patients could be given to the MRC CTU.

The application then underwent an ODR internal moderation review, and a month later, a data sharing agreement (DSA) was sent from ODR to MRC CTU. Between August and October 2019, there were ongoing discussions between the MRC CTU contracts department and the ODR. The final DSA was signed on behalf of MRC CTU on 16 October 2019 and fully executed by ODR on 15 November 2019. A further new analyst was then assigned to the project who re-reviewed the data request. This new analyst advised an update to the data censor dates, since more up-to-date data was now available from NCRAS. The updated data request was sent back to ODR for re-signing. The DSA was re-signed and the MRC CTU checked the current consent status of patients before sending participants identifiable data to NCRAS on 23 December 2019. The one-off data extracts were successfully received at the MRC CTU on 06 February 2020. This 6-week interval before data receipt was due to NCRAS rewriting their standard filters to provide C44 (non-melanoma skin cancer) — a code that is not usually supplied but needed for this trial. In total, this application, excluding the planning and preparatory work, took approximately 13 months from submission of the application to receiving the data.

PATCH

The PATCH trial opened to recruitment in 2006 as a phase II feasibility trial, developing into a phase III RCT in 2013. The trial was not initiated with the use of EHR in mind but there was a statement included in the

consent form to potentially allow information to be sought from the national registries in the future:

> I agree that my details including my full name can be given to the MRC such that long-term follow-up information from the NHS Information Centre and the NHS Central Register or any applicable NHS information system.

With the assumption of valid consent for the use of EHR data, a methodological sub-study was devised to triangulate cardiovascular event data between HES, NICOR and trial data. There was an initial scoping of the project in 2014 with NICOR and HSC-IC advising data linkage before comparison at the MRC CTU (Fig. 5). During the initial conversations with NICOR and HSC-IC, the organisations stated that the consent statement was insufficient to acquire linked data from these two sources without first gaining approval from the Confidentiality Advisory Group (CAG). In 2016, the process to submit a CAG application was started. Several months of delays followed due to difficulty in acquiring the appropriate IG documentation for PATCH. CAG require detailed IG documentation for both the trial but also in this case from NICOR and NHS Digital (formerly HSC-IC until 2016). There were difficulties in identifying the appropriate person for this information within NHS Digital, taking most of 2016 to achieve (note: at this time, case officers were not assigned until after the application was formally submitted). During 2016, an alternative method of data access was explored via NCRAS, but as no cancer data was being sought, this option was deemed unviable. Consequently, in 2017, the project was put on hold.

In October 2018, the MRC CTU re-engaged with NICOR (which had moved to Barts Health NHS Trust following a European Union tender process) and NHS Digital. There were additional complexities for obtaining CAG approval as the PATCH trial at the time was in the process of changing sponsor and therefore the CAG application could not be approved.

As the explicit wording on the consent form was the main issue preventing access to the data, the MRC CTU asked the MRC Regulatory Support Centre for further guidance. They felt that the consent wording was sufficient. NICOR subsequently agreed that, if their data was not sent to NHS Digital for linkage, then CAG approval was not necessary. Therefore a further application was submitted and sent to NICOR for review (Fig. 3). NICOR's review was completed in May 2019. The application was then submitted to HQIP by NICOR. The application was reviewed in June and amendments were

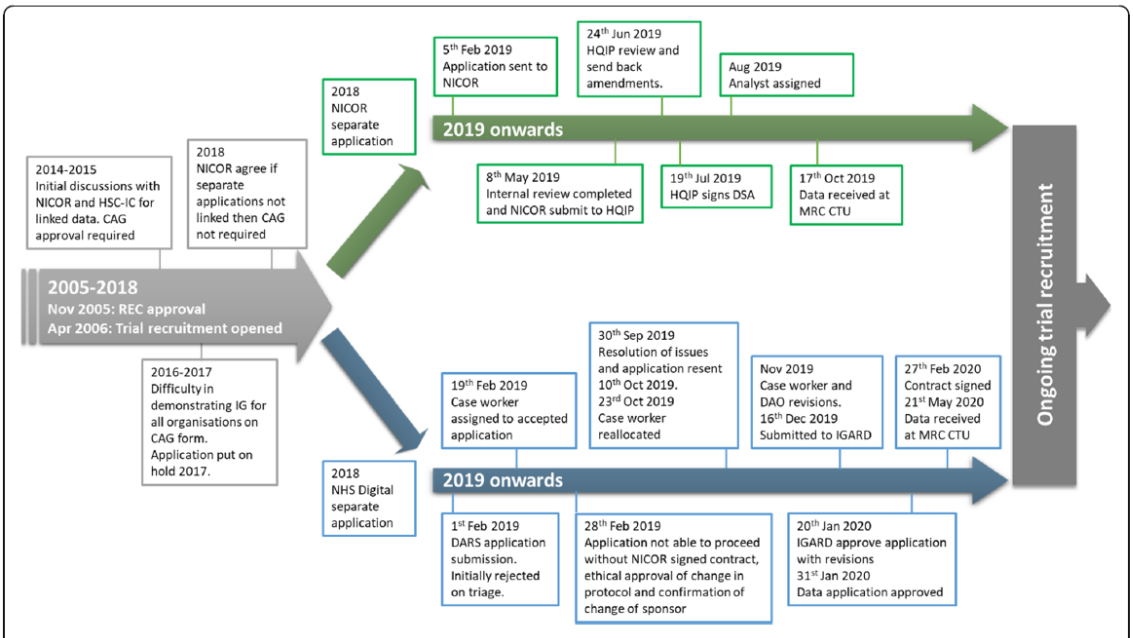


Fig. 5 Flow diagram of the PATCH joint application to NHS Digital and National Institute for Cardiovascular Outcomes Research (NICOR) and subsequently handled as separate applications in 2018. (Please note that timeline is not proportional) REC approval for PATCH November 2005. Recruitment opened in April 2006 and is ongoing. CAG, Confidentiality Advisory Group; DAO, data approvals officer; DARS, Data Access Request Service; HQIP, Health Quality Improvement Partnership; HSC-IC, Health and Social Care Information Centre; IGARD, Independent Group Advising on the Release of Data; REC, Research Ethics Committee

returned to MRC CTU. HQIP issued a signed DSA on 19 July 2019, and a NICOR analyst was assigned. The analyst continued discussions with the MRC CTU on data extraction, and a one-off data extract was received at the MRC CTU on 17 October 2019.

As with NICOR, NHS Digital was re-engaged in October 2018, and it took several weeks to allow access to the DARS online system due to technical difficulties with the DARS system (Fig. 5). A new DARS application was submitted in February 2019, but this was initially rejected due to issues around consent and sponsorship and not meeting the DARS checklist criteria. After a phone call to DARS and changes to the application by the MRC CTU, it was accepted and a case officer allocated. The case officer reviewed and made extensive comments with required changes. A privacy notice was created for the project and circulated to participants once it was ethically approved. NHS Digital then advised that the application could not proceed until the NICOR DSA was signed, sponsorship clarified and the new protocol for the sub-study had been ethically approved.

Sponsorship was not resolved until September 2019, and at that point, the MRC CTU re-engaged with NHS Digital. On receipt of the revised application, NHS Digital returned it to the DARS triage service and a new case officer was allocated. Over the next few months, the case officer made amendments to the application and sent it internally to the data approvals officer (DAO). The DAO asked for further changes to the application to clarify certain points and was submitted to IGARD in December 2019 for final review. IGARD approved the application in January subject to one last data specification amendment. The DSA was signed on behalf of the MRC CTU in February 2020, and the MRC CTU uploaded identifiable data to NHS Digital in March. The NHS Digital production team made data available in May and data was received at the MRC CTU on 21 May 2020. When all efforts are taken into consideration, it has taken several years to obtain data from both of these providers. However, from the most recent effort, data was received approximately 8 and 15 months after submission of formal applications to NICOR and NHS Digital respectively.

Discussion

This article describes the MRC CTU's experience of attempting to access EHR data from three English national data holders (NCRAS, NICOR and NHS Digital) for two large trials with a view to identifying shareable lessons. These data access applications were chosen as they were both for methodological studies embedded within RCTs looking at the appropriateness of EHR data to be used in trial follow-up with the important juxtaposition of where data access is planned versus being a later addition. The aim was to improve the knowledge and experience of gaining access to these datasets and to assess the accuracy of nationally held EHR data compared to data manually collected as part of conventional trial-specific follow-up. Our experience was challenging and took many person hours over 8 to 15 months from formally submitting an application to receiving the data.

There are limitations to this paper as this is specific to English national data holders and other countries may not have the same application issues or comparable registry data quality. This is also an experience paper from one clinical trials unit, and the difficulties we had in acquiring the data may potentially be unique. The nature of the trials, the infrastructure within this specific trials unit, the introducing of significant data protection legislation (GDPR; May 2018) during the period that provide new requirements, and the relative infrequency of our applications could be factors in the delays encountered. The process of applying for data for the PATCH trial started more than 5 years ago but the most recent iteration of applications for data started in October 2018. However, this is not a story in isolation and there have been other publications demonstrating similar problems [7, 9, 20, 21]. At present, the application process for each of these datasets is too complicated and discourages researchers from using this invaluable data. A recent survey of the cancer research community, conducted by the National Cancer Research Institute, found that less than half were successful in accessing data from the national datasets and, when asked what would help most, the majority answered 'support through data access process' and 'improving timelines for the application approval' [22]. The difficulty of accessing this data may be why so few clinical trials have used national datasets to enrich or replace data collected via conventional case report forms [8].

From a clinical trialist perspective, several lessons have been learnt about the process of applying for and obtaining EHR data. Firstly, it is extremely challenging to acquire data for an actively recruiting trial that had not planned this acquisition in advance. The main issue for the PATCH trial application was the wording in the trial protocol, consent forms and patient information sheets were not initially designed for the sub-study when the application process was started. Although the wording followed current recommendations when first written, information governance procedures and regulations evolved. In contrast, the Add-Aspirin trial had a good foundation due to prior preparation work before the application process began which meant fewer amendments were needed due to new data laws. Clinical trials units need to work closely with registries and data holders to establish the most efficient methods to obtain and access EHR data; this could include clear guidance on the

optimal timing of data requests (such as at trial initiation) and accessible, transparent cost structures to allow trialists to obtain sufficient funding for repeated data access through the lifetime of a trial.

Secondly, all clinical trials units need appropriate infrastructure to have the high level of data security needed for storing EHR data, and evidenced through a completed and endorsed Data Security and Protection Toolkit. An example includes the formation of 'Trusted Research Environments' which allow a cyber-secure virtual location where identifiable data cannot be removed and only verified researchers can access depending on IG training and specified parameters. Such infrastructure is complicated and costly taking considerable time to set up and to manage going forward. Once the required infrastructure is established, then the data security and IG controls should be valid for any national dataset. The connectivity of these datasets is also an issue, with separate applications having to be completed to several organisations/countries within the UK which takes a considerable amount of time and money. One solution would be a 'passport' system for data access to allow an institution that has demonstrated appropriate data security and IG controls to fast track the process. Another solution would be to link more datasets and allow only one application for both. There are new initiatives ongoing with examples of collaboration such as VICORI which links between NICOR and NCRAS data [23].

Lastly, the applicant also needs experience in how to answer the questions in the forms to stand up to the scrutiny of the data controllers' checks. These assessments are appropriate but, without prior knowledge, applications are often rejected due to wording rather than due to the nature of their request. This could make it difficult for clinical trials units that only apply occasionally since key knowledge may be lost inducing repetitive errors again, or the team is unaware of how the process has changed. This lack of experience can only be helped by resources provided by the dataset organisations and more guidance through the application process by experienced case officers within those organisations.

NHS Digital and NCRAS are continuing to improve their accessibility through guidelines for the application process, seminars and videos. NHS Digital has established a clinical trials service in collaboration with Health Data Research UK, the University of Oxford, IBM and Microsoft [24]. This 'NHS DigiTrials' is in its infancy and is initially concentrating on helping new trials with the identification of potential participants and follow-up of participants during and post-trial. As part of this, it is directing its attention to helping with data access from EHR for clinical trials by increasing the speed of access and a wider range of data types available. NICOR are also striving to streamline their application process internally and with HQIP to avoid unnecessary delays for appropriate research applications. During the COVID pandemic, there has also been data sharing and routine linkage for the first time between NICOR and NHS Digital that has been used in a number of publications [25].

For routinely collected EHR to be a viable option of providing data for clinical trials, data access must take no longer than a few months; otherwise, delays cause difficulty with funding and the timeliness for reporting key outcomes. Also, the records within the databases need to be up-to-date. Some may have a reporting lag of up to a year and that limits their utility. Also, better coordination and linkage between the datasets held by separate data controllers would reduce the burden on the applicants. Health Data Research UK (HDR UK) is working with key stakeholders to improve data 'inclusivity and transparency' to push the agenda of utilisation of data for science with relevant organisations but also with the public as well. This also includes improving navigation across datasets from different data controllers, via the Health Data Research Innovation Gateway, and bringing together different data controllers under the UK Health Data Alliance [26]. This is to be consistent with their bold statement of 'Our Data, Our Society, Our Health' [20]. This will hopefully allow the right data to be given to the right people in an efficient but transparent way and provide reassurance to the general public. The accessibility is the first challenge in the use of this data but there is still concern about how appropriate the data is, given that it is not designed for clinical trials. Evaluation of the reliability, completeness and accuracy of data is needed. The analysis of the EHR data of the two methodology projects described above is ongoing and will be the subject of separate publications which will further inform the discussion around the utility of EHR in trials.

Conclusion

EHR contains a wealth of information about individual patient's health outcomes, which can be useful for clinical trials. Our experience demonstrates that data access can be a prolonged and complex process. This is compounded by the fact that multiple data sources, sometimes from different data holders, will often be required for the same project. Improving data access would be the first step to realise the potential of these datasets. Based on our experience successfully accessing datasets from NHS Digital, NCRAS and NICOR, we have identified pre-planned acquisition of data prior to trial set up is important for researchers considering the use of EHR data for their clinical trials to establish appropriate consent, legal purpose and infrastructure to comply with data security and law. Data holders and researchers are endeavouring to simplify and streamline the application process so that the potential of EHR can be realised for clinical trials.

Abbreviations
CAG: Confidentiality Advisory Group of the Health Research Authority (HRA); CTU: Clinical trial unit; DAO: Data approvals officer; DARS: Data Access Request Service; DSA: Data sharing agreement; EHR: Electronic health record; GDPR: General Data Protection Regulation (EU) 2016/679; implemented 25 May 2018; HDR UK: Health Data Research UK; HES: Hospital Episodes Statistics; HQIP: Health Quality Improvement Partnership; HSC-IC: Health and Social Care Information Centre; IG: Information governance; IGAR D: Independent Group Advising on the Release of Data; MINAP: Myocardial Ischaemia National Audit Project; MRC CTU: Medical Research Council Clinical Trials Unit at UCL; NCAP: National Cardiac Audit Programme; NCIN: National Cancer Intelligence Network; NCRAS: National Cancer Registration and Analysis Service; NDR: National Disease Registration; NHFA: National Heart Failure Audit; NICOR: National Institute for Cardiovascular Outcomes Research; ODR: Public Health England (PHE) Office for Data Release; ONS: Office for National Statistics; OPCS: Office of Population Censuses and Surveys; Privacy notice: A statement made to a data subject that describes how the organisation collects, uses, retains and discloses personal information; also known as a transparency notice; REC: Research Ethics Committee

Acknowledgements
We acknowledge and are grateful to NHS Digital Research and clinical trials team and to Mark De Belder, NICOR Operational and Methodology Group Chair, and Luke Hounsome, PhD, Analytical Programme Manager at NCRAS for their comments received on the draft manuscript.

Authors' contributions
AM conceived the manuscript and led the writing. REL, SL, JC, TD, MM and MS wrote critical sections and reviewed and agreed the final version. REL, DG, FC and MP are clinical and statistical leads for Add-Aspirin and PATCH trials and reviewed and agreed the final version of the document. The authors read and approved the final manuscript.

Funding
This work was supported by Health Data Research UK; Medical Research Council MC_UU_12023/24. The Add-Aspirin trial is being jointly funded by Cancer Research UK (grant number C471 /A15015), The National Institute for Health Research Health Technology Assessment Programme (project reference 12/01/38) and the MRC Clinical Trials Unit at UCL (MC_UU_12023/28). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. The PATCH study is funded by Cancer Research UK, grant number C471/A12443 (trial CRUK/06/001) and University College London (UCL), is now sponsored by UCL and was previously sponsored by Imperial College London.

Availability of data and materials
Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Add-Aspirin was approved by the South Central – Oxford C research ethics committee and is part of the UK National Cancer Research Network (NCRN) portfolio. PATCH was approved by the Leeds (East) Research Ethics Committee.

Consent for publication
Not applicable

Competing interests
All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare no support from any organisation for the submitted work; however, MS reports grants from Health Data Research UK, during the conduct of the study; personal fees from Lilly Oncology; personal fees from Janssen; grants and non-financial support from Astellas; grants and non-financial support from Clovis Oncology; grants and non-financial support from Janssen; grants and non-financial support from Novartis; grants and non-financial support from Pfizer; and grants and non-financial support from Sanofi-Aventis, outside the submitted work; FC reports receipt of research grants for the Add-Aspirin trial from Cancer Research UK and the National Institute of Health Research, as well as study drug provision from Bayer Pharmaceuticals. REL reports grants from Cancer Research UK; grants from UK Medical Research Council, during the conduct of the study; and personal fees from Aspirin Foundation, outside the submitted work.

Author details
[1]MRC Clinical Trials Unit at UCL, UCL, London WC1V 6LJ, UK. [2]Health Data Research UK, London, UK. [3]NHS Digital, 1 Trevelyan Square, Leeds LS1 6AE, UK. [4]Medical Statistics, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK.

Received: 6 October 2020 Accepted: 24 April 2021
Published online: 10 May 2021

References

1. Lauer MS, D'Agostino RB. The randomized registry trial — the next disruptive technology in clinical research? New Engl J Med. 2013;369(17):1579–81. https://doi.org/10.1056/NEJMp1310102.
2. Mc Cord KA, Al-Shahi Salman R, Treweek S, Gardner H, Strech D, Whiteley W, et al. Routinely collected data for randomized trials: promises, barriers, and implications. Trials. 2018;19(1):29. https://doi.org/10.1186/s13063-017-2394-5.
3. Appleyard SE, Gilbert DC. Innovative solutions for clinical trial follow-up: adding value from nationally held UK data. Clin Oncol. 2017;29(12):789–95. https://doi.org/10.1016/j.clon.2017.10.003.
4. Lagerqvist B, Fröbert O, Olivecrona GK, Gudnason T, Maeng M, Alström P, et al. Outcomes 1 year after thrombus aspiration for myocardial infarction. New Engl J Med. 2014;371(12):1111–20. https://doi.org/10.1056/NEJMoa1405707.
5. Shore BJ, Nasreddine AY, Kocher MS. Overcoming the funding challenge: the cost of randomized controlled trials in the next decade. JBJS. 2012;94(Supplement_1):101–6.
6. McCord K, Hemkens L. Using electronic health records for clinical trials: where do we stand and where can we go? Can Med Assoc J. 2019;191(5):E128–E33. https://doi.org/10.1503/cmaj.180841.
7. Lugg-Widger F, Angel L, Cannings-John R, Hood K, Hughes K, Moody G, et al. Challenges in accessing routinely collected data from multiple providers in the UK for primary studies: managing the morass. Int J Popul Data Sci. 2018;3(3):1–14.
8. Lensen S, Macnair A, Love SB, Yorke-Edwards V, Noor NM, Martyn M, et al. Access to routinely collected health data for clinical trials – review of successful data requests to UK registries. Trials. 2020;21(1):398. https://doi.org/10.1186/s13063-020-04329-8.
9. Peden CJ, Stephens T, Martin G, Kahan BC, Thomson A, Rivett K, et al. Effectiveness of a national quality improvement programme to improve survival after emergency abdominal surgery (EPOCH): a stepped-wedge cluster-randomised trial. Lancet. 2019;393(10187):2213–21. https://doi.org/10.1016/S0140-6736(18)32521-2.
10. Coyle C, Cafferty FH, Rowley S, MacKenzie M, Berkman L, Gupta S, et al. ADD-ASPIRIN: a phase III, double-blind, placebo controlled, randomised trial assessing the effects of aspirin on disease recurrence and survival after primary therapy in common non-metastatic solid tumours. Contemp Clin Trials. 2016;51:56–64. https://doi.org/10.1016/j.cct.2016.10.004.
11. Public Health England. Guidance National Cancer Registration and Analysis Service 2020 [Available from: https://www.gov.uk/guidance/national-cancer-registration-and-analysis-service-ncras. Accessed 19/02/2020.
12. Langley RE, Cafferty FH, Alhasso AA, Rosen SD, Sundaram SK, Freeman SC, et al. Cardiovascular outcomes in patients with locally advanced and metastatic prostate cancer treated with luteinising-hormone-releasing-hormone agonists or transdermal oestrogen: the randomised, phase 2 MRC PATCH trial (PR09). Lancet Oncol. 2013;14(4):306–16. https://doi.org/10.1016/S1470-2045(13)70025-1.
13. Public Health England. National Cancer Intelligence Network (NCIN): 30 + years of cancer intelligence - challenges of technologies of the time [Available from: http://www.ncin.org.uk/home. Accessed 09/08/2019.
14. Public Health England. Guidance accessing PHE data through the Office for Data Release 2020 [Available from: https://www.gov.uk/government/publications/accessing-public-health-england-data/about-the-phe-odr-and-accessing-data. Accessed 19/02/2020.

15. Gov.uk. HSCIC changing its name to NHS Digital 2016 [Available from: https://www.gov.uk/government/news/hscic-changing-its-name-to-nhs-digital. Accessed 19/02/2020.
16. Boyd A. Understanding Hospital Episode Statistics (HES). London, UK: CLOSER; 2017.
17. NHS Digital. Data Access Request Service (DARS): process 2019 [Available from: https://digital.nhs.uk/services/data-access-request-service-dars/data-access-request-service-dars-process. Accessed 20/02/2020.
18. NICOR. NICOR 2020 [Available from: https://www.nicor.org.uk/. Accessed 20/02/2020.
19. McKay AJ, Jones AP, Gamble CL, et al. Use of routinely collected data in a UK cohort of publicly funded randomised clinical trials. F1000Research. 2020;9:323.
20. Ford E, Boyd A, Bowles JKF, Havard A, Aldridge RW, Curcin V, et al. Our data, our society, our health: a vision for inclusive and transparent health data science in the United Kingdom and beyond. Learning Health Systems. 2019; 3(3):e10191. https://doi.org/10.1002/lrh2.10191.
21. Dattani N, Hardelid P, Davey J, Gilbert R. Accessing electronic administrative health data for research takes time. Arch Dis Childhood. 2013;98(5):391–2. https://doi.org/10.1136/archdischild-2013-303730.
22. National Cancer Research Institute. The researchers' experience when attempting to access health data for research 2020 [Available from: https://www.ncri.org.uk/ncri-blog/accessing-health-data-for-research/. Accessed 28/02/2020.
23. Public Health England. Current analytical partnerships involving the National Cancer Registration and Analysis Service 2019 [Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/787750/Current_analytical_partnerships_involving_NCRAS.pdf. Accessed 01/05/2020.
24. NHS Digital. NHS DigiTrials 2020 [Available from: https://digital.nhs.uk/services/nhs-digitrials. Accessed 01/06/2020.
25. Mohamed MO, Gale CP, Kontopantelis E, Doran T, de Belder M, Asaria M, et al. Sex-differences in mortality rates and underlying conditions for COVID-19 deaths in England and Wales. Mayo Clinic Proceedings. 2020;95(10): 2110–24. https://doi.org/10.1016/j.mayocp.2020.07.009.
26. Health Data Research UK. Health Data Research Innovation Gateway 2020 [Available from: https://www.healthdatagateway.org/. Accessed 01/06/2020.

Publisher's Note

Appendix C: Summary of contracted data summary transferred from PHE with ODR

Annex A: Data Transfer Form

ODR reference:	ODR1718_261
Application title:	A phase III, double-blind, placebo-controlled, randomised trial assessing the effects of aspirin on disease recurrence and survival after primary therapy in common non-metastatic solid tumours

Subject to the terms and conditions of the contract ODR1718_261, PHE agrees to transfer the Data as identified below:

Detailed Data Specification

Cohort inclusion/exclusion criteria:

Inclusion

- Participants in the Add-Aspirin trial (EUDRACT # 2013-004398-28) randomised between 01 December 2015 and 30 September 2018
- Extant consent
- Resident in England

Tumour and event level requirements

- For each linked participant, all tumours (C00-D48) diagnosed from 01 December 2012 to 30 June 2018 will be provided
- HES episodes will be limited to records from 31 days prior to diagnosis_best of the earliest first primary of the following solid tumour C50, C18-C20, C15, C16, C61 up to 31 March 2019 (most available data at time of release)
- Index and censor dates associated with all other tables are documented with the table
- CDF restrictions will not be applied

Table 1: Patient identifiers to be provided by UCL to PHE for data linkage to the cancer registry in accordance with the cohort inclusion criteria.

Unique trial-specific ID
NHS number
DOB

Table 2: AV_patient data to be released to the data recipient by NCRAS for all registry-matched trial participants

Unique trial-specific ID
VITALSTATUS
VITALSTATUSDATE
DEATHCAUSECODE_1A
DEATHCAUSECODE_1B
DEATHCAUSECODE_1C
DEATHCAUSECODE_2
DEATHCAUSECODE_UNDERLYING
DEATHLOCATIONDESC

POSTMORTEM
TUMOURCOUNT
BIGTUMOURCOUNT

Table 3: AV_tumour data to be released to the data recipient by NCRAS for all registry-matched trial participants. All registerable tumours (C00-D48) diagnosed between 01 December 2012 to 30 June 2018 will be provided.

Unique trial-specific ID
TUMOURID (pseudonymised)
DIAGNOSISDATEBEST
BASISOFDIAGNOSIS
DCO
SITE_ICD10_O2
SITE_CODED
CODING_SYSTEM_DESC
MORPH_ICD10_O2
BEHAVIOUR_ICD10_O2
HISTOLOGY_CODED
GRADE
TUMOURSIZE
nodes_involved_new
LATERALITY
ER_STATUS
HER2_STATUS
DUKES
GLEASON_COMBINED
STAGE_IMG
T_PATH
N_PATH
M_PATH
T_BEST
N_BEST
M_BEST
STAGE_BEST
DATE_FIRST_SURGERY
TRUSTCODE_FIRST_SURGERY

Table 4: AV_treatment data to be released to the data recipient by NCRAS for event linked to registerable tumours (C00-D48) diagnosed between 01 December 2012 to 30 June 2018

EVENTID (pseudonymised)
TUMOURID (pseudonymised)
Unique trial-specific ID
EVENTCODE
EVENTDESC
EVENTDATE
OPCS4_CODE
RADIOCODE

RADIODESC
IMAGINGCODE
IMAGINGDESC
IMAGINGSITE

Table 5: HES Inpatient (admitted care) data to be released to the data recipient by NCRAS in accordance with the inclusion criteria. HES episodes will be indexed 31 days prior to diagnosis_best of the earliest first primary solid tumour C50, C18-C20, C15, C16, C61 and censored to 31 March 2019.

Unique trial-specific ID
admidate
admimeth
admisorc
disdate
disdest
dismeth
diag_nn
opertn_nn
opdate_nn
mainspef
tretspef

Table 6: HES OP data to be released to the data recipient by NCRAS in accordance with the inclusion criteria. HES episodes will be indexed 31 days prior to diagnosis_best of the earliest first primary solid tumour C50, C18-C20, C15, C16, C61 and censored to 31 March 2019.

Unique trial-specific ID
apptdate
atentype
attended
firstatt
stafftyp
outcome
priority
diag_nn
opertn_nn
operstat
mainspef
tretspef

Table 7: HES A&E data to be released to the data recipient by NCRAS in accordance with the inclusion criteria. HES episodes will be indexed 31 days prior to diagnosis_best of the earliest first primary solid tumour C50, C18-C20, C15, C16, C61 and censored to 31 March 2019.

Unique trial-specific ID
aearrivalmode
aeattendcat

Appendix D: Upper GI, Lower GI and Intracranial Haemorrhage ICD-10 codes

Upper GI Haemorrhage ICD-10 codes

ICD-10 Code	Event description
I850	Oesophageal varices with bleeding
K226	Gastro-oesophageal laceration-bleed syndrome
K250	Gastric ulcer - Acute with bleed
K251	Gastric ulcer - Acute with perforation
K252	Gastric ulcer - Acute with both bleed and perforation
K254	Gastric ulcer - Chronic or unspecified with bleed
K255	Gastric ulcer - Chronic or unspecified with perforation
K256	Gastric ulcer - Chronic or unspecified with both bleed and perforation
K260	Duodenal ulcer - Acute with bleed
K261	Duodenal ulcer - Acute with perforation
K262	Duodenal ulcer - Acute with both bleed and perforation
K264	Duodenal ulcer - Chronic or unspecified with bleed
K265	Duodenal ulcer - Chronic or unspecified with perforation

K266	Duodenal ulcer - Chronic or unspecified with both bleed and perforation
K270	Peptic ulcer, site unspecified - Acute with bleed
K271	Peptic ulcer, site unspecified - Acute with perforation
K272	Peptic ulcer, site unspecified - Acute with both bleed and perforation
K274	Peptic ulcer, site unspecified - Chronic or unspecified with bleed
K275	Peptic ulcer, site unspecified - Chronic or unspecified with perforation
K276	Peptic ulcer, site unspecified - Chronic or unspecified with both bleed and perforation
K280	Gastrojejunal ulcer - Acute with bleed
K281	Gastrojejunal ulcer - Acute with perforation
K282	Gastrojejunal ulcer - Acute with both bleed and perforation
K284	Gastrojejunal ulcer - Chronic or unspecified with bleed
K285	Gastrojejunal ulcer - Chronic or unspecified with perforation
K286	Gastrojejunal ulcer - Chronic or unspecified with both bleed and perforation
K290	Acute haemorrhagic gastritis
K920	Haematemesis

K921	Melaena
K922	Gastrointestinal bleed, unspecified

Lower GI Haemorrhage ICD-10 codes

ICD-10 Code	Event description
K552	Angiodysplasia of colon with bleeding
K625	Bleeding of anus and rectum
K922	Gastrointestinal bleeding, unspecified

Intracranial Haemorrhage ICD-10 codes

ICD-10 Code	Event description
I60	Subarachnoid bleed
I600	Subarachnoid bleed from carotid siphon and bifurcation
I601	Subarachnoid bleed from middle cerebral artery
I602	Subarachnoid bleed from anterior communicating artery
I603	Subarachnoid bleed from posterior communicating artery
I604	Subarachnoid bleed from basilar artery
I605	Subarachnoid bleed from vertebral artery
I606	Subarachnoid bleed from other intracranial arteries
I607	Subarachnoid bleed from intracranial artery, unspecified
I608	Other subarachnoid bleed
I609	Subarachnoid bleed, unspecified
I61	Intracerebral bleed
I610	Intracerebral bleed in hemisphere, subcortical
I611	Intracerebral bleed in hemisphere, cortical
I612	Intracerebral bleed in hemisphere, unspecified
I613	Intracerebral bleed in brain stem
I614	Intracerebral bleed in cerebellum
I615	Intracerebral bleed, intraventricular

I616	Intracerebral bleed, multiple localized
I618	Other intracerebral bleed
I619	Intracerebral bleed, unspecified
I62	Other nontraumatic intracranial bleed
I620	Subdural bleed (acute)(nontraumatic)
I621	Nontraumatic extradural bleed
I629	Intracranial bleed (nontraumatic), unspecified
S065	Traumatic subdural bleed
S0650	Traumatic subdural bleed - without open intracranial wound
S0651	Traumatic subdural bleed - with open intracranial wound
S066	Traumatic subarachnoid bleed
S0660	Traumatic subarachnoid bleed - without open intracranial wound
S0661	Traumatic subarachnoid bleed - with open intracranial wound