BMJ Quality & Safety

Variations in GPs' Decisions to Investigate Suspected Lung Cancer: A Factorial Experiment Using Multimedia Vignettes

Journal:	*BMJ Quality & Safety*
Manuscript ID	bmjqs-2016-005679.R2
Article Type:	Original Research
Keywords:	Diagnostic errors, General practice, Primary care, Simulation, Health services research

ABSTRACT

INTRODUCTION: Lung cancer survival is low and comparatively poor in the UK. Patients with symptoms suggestive of lung cancer commonly consult primary care but it is unclear how general practitioners (GPs) distinguish which patients require further investigation. This study examined how patients' clinical and socio-demographic characteristics influence GPs' decisions to initiate lung cancer investigations.

METHODS: A factorial experiment was conducted amongst a national sample of 227 English GPs using vignettes presented as simulated consultations. A multimedia interactive website simulated key features of consultations using actors ('patients'). GP participants made management decisions online for six 'patients', whose socio-demographic characteristics systematically varied across three levels of cancer risk. In low-risk vignettes, investigation (i.e. chest X-ray ordered, computerised tomography scan or respiratory consultant referral) was not indicated; in medium-risk, investigation could be appropriate; in high-risk vignettes, investigation was definitely indicated. Each 'patient' had two lung cancer-related symptoms; one volunteered and another elicited if GPs asked. Variations in investigation likelihood were examined using multilevel logistic regression.

RESULTS: GPs decided to investigate lung cancer in 74% (1000/1348) of vignettes. Investigation likelihood did not increase with cancer risk. Investigations were more likely when GPs requested information on symptoms that 'patients' had but did not volunteer (adjusted odds ratio (AOR)=3.18; 95%CI 2.27-4.70). However GPs omitted to seek this information in 42% (570/1348) of cases. GPs were less likely to investigate older than younger 'patients' (AOR=0.52 95%CI 0.39-0.7) and Black 'patients' than White (AOR=0.68; 95%CI 0.48-0.95).

CONCLUSIONS: GPs were not more likely to investigate 'patients' with high than low-risk cancer symptoms. Furthermore, they did not investigate everyone with the same symptoms equally. Insufficient data gathering could be responsible for missed opportunities in diagnosis.

INTRODUCTION

Lung cancer, the most common cancer worldwide, has comparatively poor survival in the UK.[1] Most lung cancer patients first present to primary care but diagnostic delays are well documented: lung cancer patients have more consultations in primary care before investigation than many other cancers.[2] In addition, whilst intervals from presentation to diagnosis have reduced for other common cancers over time, they remain unchanged for lung cancer.[3] It has been suggested that missed opportunities for lung cancer diagnosis in primary care may contribute to poor lung cancer survival.[4]

Primary care physicians, referred to throughout this paper as general practitioners (GPs), have direct access to lung cancer diagnostic tools including chest X-ray. GPs may not consider lung cancer as a differential diagnosis because patients with lung cancer commonly present in primary care with non-specific symptoms that are more often due to benign causes.[5] Non-specific symptoms and rare disease occurrence therefore present diagnostic difficulty for GPs.[6] Reducing diagnostic delays requires an understanding of how GPs decide which patients with common, non-specific symptoms to investigate for lung cancer. Not only is it unclear how GPs decide who requires further investigation by chest X-ray or by specialist referral, but inequalities by patient age, gender and socioeconomic circumstances have been identified in retrospective analyses of routine data.[1,2,7,8] Most previous research has examined the diagnostic process using retrospective data in cancer patients only,[5] thus missing a key dimension, i.e. how GPs decide which patients with symptoms do not require investigation.

Examining decision making in a standardised way in clinical practice presents substantial methodological challenges.[9,10] Direct observation of real physician-patient encounters offers no opportunity to control patients' clinical and socio-demographic characteristics, and so requires observation of very large numbers of consultations to obtain the necessary numbers in specific risk or demographic categories. The use of fictional patient profiles (vignettes) can provide a valid and efficient approach to examining clinician behaviour,[11] and studies have already produced useful insights into sources of error in clinicians' decision making processes, due to both patient factors (e.g. symptom characteristics)[12] and physician factors (e.g. cognitive biases).[12,13] As Blumenthal-Barby and others recognise, however, there are limits to the applicability of written vignettes and other vignette designs that do not simulate key features of real consultations.[14] In particular, when vignettes offer little or no opportunity for physicians to seek information from or about the vignette patient, they can inappropriately frame the decision for the physician by cueing what they should notice about the patient or by

offering participants only a limited selection of response options. This risks priming participating physicians to consider certain actions, and biasing their responses.

In this vignette study, we therefore sought to simulate key features of consultations. We designed a website using interactive multimedia vignettes with videos of actor 'patients', which enabled participating GPs to ask questions in their own words and receive real-time responses. We used this intervention in a factorial randomised experimental study to examine GPs' decisions to initiate lung cancer investigation across different combinations of patient clinical and socio-demographic characteristics.

METHODS

Design

We constructed 36 simulated consultations comprising video vignettes of actor 'patients' and comprehensive clinical information, including previous medical history, co-morbidities and examination findings, and socio-demographic characteristics,. The symptomatic information provided adhered to material in the latest available National Institute for Health and Care Excellence (NICE) referral guidelines for suspected cancer (published in 2005),[15] with cancer risk based on data from the CAPER case-control study.[16] Each consultation was designed to take participating GPs approximately 10 minutes to complete so that it mirrored the length of a 'real' clinical encounter in primary care in the UK National Health Service.

At the start of each 'consultation', a video was shown where the actor 'patient' volunteered a description of their presenting symptom. Participants could then elicit further information in real-time on the presenting symptom, other symptoms, and risk factors by typing in questions, to which they received the 'patient's' video response. They could also, if they wished, click on a drop-down menu to obtain information on behavioural and familial risk factors, previous medical history, family history, socio-demographic information and examination findings.(**Figure 1**) A demonstration is available at: www.ucl.ac.uk/stream/media/swatch?v=c22f1a2b58b8

<<FIGURE 1>>

We applied a factorial experimental design, where GPs undertook one consultation from each of six clinical profiles across three lung cancer risk levels (Table 1); no GP saw the same actor twice. Within these constraints, allocation of GPs to vignettes was random. This achieved approximate balance of patient characteristics by clinical profile, gender, ethnicity, and socioeconomic circumstances. The study protocol is available at: http://www.ucl.ac.uk/dahr/research-pages/gp_study

Recruitment and participation

Qualified GPs and registrars nearing the end of their specialist GP training were invited through nine Primary Care Research Networks across England in 2012 and 2013 to participate in a study of decision making (without explicit reference to lung cancer). Those that returned an expression of interest were sent further information. For GPs that wished to take part, their internet browsers were checked for compatibility with the study software.

GP participants were first trained to use the on-line simulated consultations. This was done using a web based video in advance of the study with access to support from the research team during or between study consultations. Each participating GP used the study website to 'consult' with six 'patients' and at the end of the 'consultation', entered their management plan. GPs also completed a brief questionnaire about their practice characteristics and years since qualifying.

Application development

The application's development followed the steps recommended by Adler et al[17] for developing simulations:

1. Case concept: developing the vignette design and content
2. Review and Revision by Content Experts
3. Outline and Flow Development: A typical online consultation in the study
4. Translation of content into simulation platform: vignette interactive website
5. Pilot testing and revisions

A detailed description of each step is given in supplementary file S1. In brief, the structure of the factorial experiment required 36 unique vignette combinations to cover the four experimental factors: known to be associated with variation in lung cancer survival, but whose effect on inequalities in GPs' rates of referral for investigation or to secondary care is uncertain[8]:

- **Ethnicity**: three variations (White, Black Caribbean, South Asian)
- **Gender**: two variations (male, female)
- **Socioeconomic circumstances**: two variations (advantaged or disadvantaged)
- **Clinical risk of lung cancer**: three variations (low, medium and high-risk), with two profiles for each level of risk. Age was not included as a separate experimental factor but was instead incorporated into profiles because older age increases the risk of cancer associated with most symptom combinations.[16] We constructed six clinical

profiles, two for each risk level using different combinations of symptoms, age, and smoking status.(Table 1) The positive predictive values (PPV) of lung cancer were drawn from PPVs generated by analysis of symptom combinations in the CAPER case-control dataset and interpretation of these symptoms and their characteristics informed by the latest available NICE guidance on investigation of suspected cancer.[15, 16] (described further in supplementary data)

To maximise the clinical authenticity of the cases, GPs specializing in cancer diagnosis and non-academic GPs reviewed the proposed vignettes. The website content and functionality were also informed by patient representatives' comments. For example, these influenced the types of responses 'patients' provided, because patient representatives corroborated previous research that patients may well not disclose certain symptoms with their doctors without being directly asked about them.[18]

The translation of content into the online study application website (virtual patient application) required filming actors portraying patients, creating and populating the website with that content. The website architecture and application software was produced by Athenaeum Educational Technologies. It involved the development of a bespoke system using natural language processing principles to recognise GPs' free-text questions and play a video clip in response (see Doan et al 2014 for an explanation of the principles).[19] This system was underpinned by databases on symptoms or risk factors and the features those symptoms (e.g. what exacerbates or relieves the symptom or how long it has been present).

Table 1. Components of the six different clinical profiles by risk level

Clinical Profile	Information volunteered by 'patient' or available onscreen			Information only available if participant GPs asked		Positive Predictive Value (PPV) of lung cancer	Other relevant information
	Age range	Smoking status	Symptom 1	Symptom 2	Duration		
Low risk: Expected action = no active investigation (safety netting appropriate)							
1	Younger (Late fifties)	Non smoker	Breathlessness	Fatigue	1-2 weeks	0.40%	Patient has swollen ankles, possibly due to heart failure
2	Younger (Late fifties)	Smoker	Chest pain	Cough	1-2 weeks	1.10%	
Medium risk: Expected action = either investigation (e.g. order chest x-ray) or safety netting							
3	Older (Late seventies)	Smoker	Chest pain	Cough	Uncertain (approx 3 weeks)	1.70%	
4	Older (Late seventies)	Non-smoker	Cough	Appetite loss	Uncertain (approx 3 weeks)	2.50%	
High risk: Expected action = lung cancer investigation							
5	Younger (Late fifties)	Smoker	Breathlessness	Fatigue	>5 weeks	3-4%	Chronic obstructive pulmonary disease (COPD) present
6	Older (Late seventies)	Smoker	Chest pain	Weight loss	>5 weeks	14%	

Analysis

Every action performed by GPs on the website (i.e. all the questions asked of 'patients', drop-down menus accessed, free-text entered in management plans) was captured by the study website. This information was used to measure the duration of each consultation and to generate three indicators about GPs' information requests in each consultation and the capacity of the research application to respond to these requests:

- *data sought*: average number of data items sought (questions asked or drop-down menu items accessed), by GP and by individual vignette
- *errors*: error messages displayed as a proportion of all data items sought, calculated for all consultations, consultation 1 and consultations 2-6 only, assuming that in the first consultation GPs were familiarising themselves with the application
- *key information elicited*: proportion of GPs that elicited information on the vignettes second, but unvolunteered, lung cancer symptom.

GPs also had the opportunity to provide free-text comments on any aspect of the application in an online survey after all the consultations were completed. These comments were not treated as a representative survey of all participants' experiences but were examined to provide insights into GPs' experiences of the application and their perceptions of its utility as a research tool for eliciting the decision making process.

The primary outcome was the proportion of 'patients' for whom lung cancer investigation was included in the management plan. This included ordering appropriate imaging, or referral for a specialist opinion e.g. from a respiratory consultant whether participants' management plan stated this investigation was for lung cancer or not. This outcome variable was constructed from free-text responses entered by participants in their management plan, according to pre-defined criteria. A clinician confirmed the validity of every constructed primary outcome.

Data were analysed by fitting multilevel logistic regression models using Markov Chain Monte Carlo for estimation,[20] allowing variation between participants and between vignettes within participants. This allowed for a correlation between outcomes within a given GP but independent outcomes for two vignettes viewed by different GPs. Estimation of odds ratios and 95% credible intervals was carried out using the RStan library in R version 3.0.2.[21] Significance testing was carried out using Wald tests based on the means and posterior variances of the estimates.

Variations in outcome were examined by 'patient' gender, ethnicity, socioeconomic circumstances and risk profile, an indicator variable for whether participants sought the second symptom, and GP characteristics (demographics, experience, and region). Two models were built in order to examine differences by a) clinical profile and b) by age. A supplementary analysis was conducted to examine whether findings were difficulties in obtaining information sought from the application, by including the indicator on errors as another covariate in each model. To examine selection bias, the gender and age of participating GPs and their practices' cancer referral characteristics were compared with national data.[22,23]

The required sample size was calculated on the basis that a minimum difference in investigations of 10% was considered of clinical importance and realistic given variations in cancer investigations in other studies.[24] A response from 216 participants was sought to give 1296 vignettes (i.e. each of the 36 vignettes viewed 36 times). Each risk and ethnic group would therefore be viewed 432 times, each gender and socioeconomic group 648 times. Assuming a 20% variance inflation factor for clustering of GPs/'patients', 432 in each risk and ethnic group would give 95% power to detect a difference of 10%. For differences between gender and socioeconomic groups, 648 in each group would give 85% power for a difference of 5%.

RESULTS

Sample characteristics

227 GPs completed the study. This was 76% of the 300 GPs who registered for the study and 41% of the 556 GPs in total that initially expressed an interest in taking part(see: supplementary file S2A). There were no demographic differences between registered GPs who did and did not complete the study but GP participants were younger than the national GP population and practices had higher cancer referrals than non-participating practices.(See: supplementary file S2B)

Out of 1362 vignettes, 14 (1%) were excluded due to missing participant demographic data in one GP (n=6, 0.4%), when participants asked about second symptoms but did not receive a response (n=4, 0.3%) or when participants did not enter a management plan (n=4, 0.3%).

Consultation process

GPs spent on average 13 minutes on the first consultation and 11 minutes on consultations 2-6 and sought 47 items of information per consultation (by asking text questions of the patient, looking up patient history or personal information, conducting 'examinations' or 'bedside tests'). GPs received error messages in response to an average of 4.6% of data sought for consultations 2-6 (range 4-22%).(See supplementary file, S2C)

Lung cancer investigations

Participants initiated investigations in 1000 (74%) vignettes. There was little difference in investigation between low, medium and high-risk levels (72-75%) but large variation between clinical profiles (59-86%). There were no variations by 'patient' gender or socioeconomic circumstances but there was a gradient in investigation by ethnicity, with 'patients' of Black ethnicities least and White ethnicities most likely to be investigated (71% vs 77%). (Table 2)

GPs asked for additional, relevant information about second symptoms in 778 (58%) of cases overall with marked variation by clinical profile, ranging from 48 (21%) in Profile 1 to 214 and 216 (95%) in Profiles 2 and 3. There was a significant interaction between seeking a relevant second symptom and clinical profile ($p<0.001$). 91% of GPs who discovered the presence of weight loss initiating investigation compared with just 46% who did not seek this information. In contrast, knowing 'patients' experienced fatigue did not significantly change the likelihood of investigation.(Table 3)

While obtaining second symptom information was associated with more investigation (adjusted odds ratio (AOR): 3.18 [2.27;4.70], $p<0.001$), there was still under-investigation in 'patients' with appetite or weight loss (Profiles 4 and 6) compared with 'patients' with chest pain and cough (Profile 3) (AORs: 0.25 [0.14;0.42], $p<0.001$; and 0.5 [0.29;0.91], $p=0.02$ respectively).(Table 4a) GPs were less likely to investigate older than younger 'patients' (AOR: 0.52 [0.39;0.70], $p<0.001$), and less likely to investigate 'patients' of Black compared with White ethnicities (AOR: 0.68 [0.48;0.95], $p=0.03$).(Table 4b)

Associations were similar when the variable for errors received was included. (See: supplementary file S2D)

Comments volunteered by GP participants on their experiences of the application and their perceptions of its utility as a research tool for eliciting the decision making process are summarised in S3.

Table 2. Frequency of lung cancer investigation

		Investigation		N (vignettes)
		n	%	
Total		**1000**	**74.18**	**1348**
a. By 'patient' characteristic				
Risk level	Low	339	75.00	452
	Medium	327	72.35	452
	High	334	75.23	444
Clinical profile[1]	Clinical Profile 1 *PPV=0.4% (younger; ns; 1-2w breathless [& fatigue])*	152	66.96	227
	Clinical Profile 2 *PPV=1.1% (younger, s; 1-2w chest pain [& cough])*	187	83.11	225
	Clinical Profile 3 *PPV=1.7% (older, s; ~3w chest pain [& cough])*	195	85.90	227
	Clinical Profile 4 *PPV=2.5% (older, ns; ~3w cough [& appetite loss])*	132	58.67	225
	Clinical Profile 5 *PPV=3-4% (younger, s; >5w breathless [& fatigue])*	185	82.59	224
	Clinical Profile 6 *PPV~14% (older, s; >5w chest pain [& weight loss])*	149	67.73	220
Gender	Female	489	74.09	660
	Male	511	74.27	688
Socioeconomic circumstances	Disadvantaged	508	74.49	682
	Advantaged	492	73.87	666
Ethnicity	White	369	76.56	482
	Black	306	71.50	428
	South Asian	325	74.20	438

2nd symptom elicited	No	361	63.33	570
	Yes	639	82.13	778
b. By GP participant characteristic				
GP gender	Female	425	70.48	603
	Male	573	77.12	743
GP age range	25-34 years	227	70.06	324
	35-44 years	336	72.89	461
	45-54 years	325	78.69	413
	55-64 years	102	75.00	136
	65 years or over/missing	8	66.67	12
Years since qualifying	0 to 2 years ago	120	71.43	168
	2 to 5 years ago	186	69.14	269
	5 to 10 years ago	177	73.75	240
	10 to 20 years ago	256	77.58	330
	20+ years ago	259	76.40	339
Ethnicity	White	583	73.89	789
	Black	34	80.95	42
	South Asian	296	73.63	402
	Other/missing	90	75.63	119
Region	London	365	73.44	497
	East of England	341	74.95	455

North West	131	76.16	172
West Midlands	96	72.73	132
Surrey and Sussex	41	75.93	54
Locum GP	24	66.67	36

[1] younger = late fifties; older = late seventies; s = smoker ns = non-smoker; w = weeks; [symptom] = not volunteered by patient

Table 3. Lung cancer investigation by profile according to whether GPs did or did not elicit symptom information

Clinical profile (Second symptom)[1]	Second symptom		Lung cancer investigation				
	Not elicited	Elicited	Symptom not elicited		Symptom elicited		Total
	n (%)	n (%)	n (%)	Odds Ratio [95% CI]	n (%)	Odds Ratio [95% CI]	n (%)
1 (Fatigue)	179 (78.85)	48 (21.15)	120 (66.67)	1.00 [-]	31 (65.96)	0.94 [0.43;2.09]	152 (66.96)
2 (Cough)	11 (4.89)	214 (95.11)	7 (63.64)	0.73 [0.16;3.18]	181 (84.19)	2.83 [1.82;4.40]	187 (83.11)
3 (Cough)	11 (4.85)	216 (95.15)	7 (63.64)	0.93 [0.19;4.39]	189 (87.1)	3.67 [2.13;6.30]	195 (85.90)
4 (Appetite loss)	89 (39.56)	136 (60.44)	42 (46.67)	0.38 [0.21;0.69]	91 (66.91)	0.98 [0.59;1.62]	132 (58.67)
5 (Fatigue)	168 (75.00)	56 (25.00)	136 (80.47)	2.21 [1.31;3.72]	50 (89.29)	4.59 [2.86;7.37]	185 (82.59)
6 (Weight loss)	112 (50.91)	108 (49.09)	52 (46.02)	0.36 [0.20;0.62]	99 (90.83)	5.69 [2.07;15.63]	149 (67.73)
Total	**570 (42.28)**	**778 (57.72)**	**364 (63.41)**		**641 (82.18)**		**1000 (74.18)**

[1] Clinical profile is formed from symptoms, smoking status and patient age

Table 4. Multilevel logistic regression of cancer investigation by 'patient' characteristic

a) By clinical profile		**Adjusted[1] odds ratio [95% CI]**
Clinical profile (2nd symptom)	1 (Fatigue)	0.62 [0.35; 1.10]
	2 (Cough)	0.65 [0.38; 1.15]
	3 (Cough)	1
	4 (Weight loss)	0.25 [0.14; 0.42]*
	5 (Fatigue)	1.64 [0.90; 3.11]
	6 (Appetite loss)	0.50 [0.29; 0.91]*
Ethnicity	White	1
	South Asian	0.86 [0.62; 1.20]
	Black	0.67 [0.47; 0.96]*
Second symptom elicited	No	1
	Yes	3.18 [2.27; 4.70]*
b) By age		
Age	Younger (Late fifties)	1
	Older (Late seventies)	0.52 [0.39; 0.70]*
Ethnicity	White	1
	South Asian	0.88 [0.63; 1.27]
	Black	0.68 [0.48; 0.95]*
Smoking status	Non smoker	1
	Smoker	2.24 [1.64; 3.02]*
Second symptom elicited	No	1
	Yes	2.83 [2.09; 3.83]*

[1] adjusted for all other factors associated ($p<0.1$) with investigation in univariate analysis (i.e. 'patient' profile and ethnicity, GP gender and age), and whether second symptom was elicited

[2] adjusted for 'patient' profile, ethnicity, GP gender and age and whether second symptom was elicited

* significant at $p\leq0.05$

DISCUSSION

Summary

In this factorial experiment using vignettes in simulated consultations, GPs' decisions to investigate lung cancer was influenced by whether they sought out additional, relevant clinical information about the presence of common symptoms. Even when participating GPs elicited sufficient information about symptoms, inequalities by age and ethnicity in investigation decisions remained.

Comparisons with existing literature

Our data were collected during 2012-2013 and our finding that GPs investigated a high proportion (72-75%) of cases is in line with literature from 2013.[24] However it is higher than might have been expected if GPs were following the latest national guidance for suspected cancer investigation available during the study period.[15] Participants may have proposed more tests for vignette 'patients' than they would in reality because they were not subject to the resource constraints of clinical practice or may have ordered X-rays primarily to investigate diagnoses other than cancer. Alternatively, they may have been aware of and responding to epidemiological evidence, presumed patient preferences, and policy published since the 2005 NICE guidance, all of which support a lower threshold for cancer investigation.[25-28] Indeed, updated NICE guidance on referral of suspected cancer, published in 2015 (after our data were collected), include a substantially lower investigation threshold than that recommended in their earlier guideline,[29] such that all our vignettes would now suggest investigation.

We found that in 42% of cases, GPs did not seek additional information that would help to make an informed decision regarding referral and that was available on request. This accords to some extent with international studies of missed opportunities in cancer diagnosis.[30-31] In the UK, the updated NICE guidance explicitly recognises that patients with combinations of common symptoms may be more likely to have lung cancer than patients with any one of these symptoms alone,[29,32] but patients may not volunteer all the symptoms they experience in consultations, perhaps due to real or perceived time constraints in the consultation.[32] The importance of data gathering for reaching a timely diagnosis was highlighted in the recent Institute of Medicine Report into improving diagnosis in health care.[33] Zwaan et al's study of breathlessness using expert review of medical records found evidence of inappropriately

selective information gathering in a third of cases with some evidence that diagnostic error and patient harm occurred in a proportion of these cases.[34] Our study extends the field by providing objective evidence of non-clinical variations in data gathering by physicians in a large vignette study and demonstrates associations between gathering sufficient data and appropriate decision making.

We also found that the effect of eliciting this second symptom on decision making varied by symptom. It made little difference whether participants knew that patients had a cough or fatigue, but made significant difference to decision making if participants knew of appetite and weight loss. For weight loss in particular (a key question when clinicians are considering whether cancer is a possible diagnosis), in 91% of cases where GP participants had elicited information about weight loss, they initiated investigation, compared with just 46% where GPs were unaware the patient had lost weight. It is important to acknowledge that neither in real life nor in the vignettes are the factors (symptom, age and smoking) that constituted each profile independent of one another. Therefore whilst we contend the results are interpretable and reliable, they are not as definitive as a randomised controlled trial results so this finding has to be treated with some caution. However, the finding accords with Kostopoulou et al's recent 'think aloud' study which suggests that when physicians have an idea of cancer early in the consultation, they ask pertinent questions and initiate appropriate investigations to ensure a cancer diagnosis is reached.[35] Therefore, it still seems likely that routinely questioning patients with ongoing respiratory symptoms about weight loss would expedite the diagnosis of some lung cancers.

Our finding that GPs were less likely to investigate older 'patients' is consistent with several observational studies of primary care cancer referral and investigation.[36,37] Scott et al's Model of Pathways to Treatment proposes that as patients grow older, they are increasingly likely to attribute bodily changes to normal ageing processes than to disease.[38] If clinicians also apply this 'normal ageing' heuristic, it may explain why GPs in this study were less likely to investigate older patients, despite knowing their symptoms. In contrast, patient experience survey data indicate more referral delays in younger (aged 55-64 years) than older patients (over 75 years). However survey data may be biased if older patients (with lower overall survival) were underrepresented because they had died or were too ill to participate in the survey (which was undertaken 6-12 months after diagnosis).[2]

We also found smaller ethnic variations in GPs' investigation behaviour, with fewer investigations initiated in Black (and to some extent) South Asian 'patients' than White. This is consistent with survey data where non-White cancer patients report more referral delays than White patients.[2] One possible explanation is that GPs were less ready to consider a lung cancer diagnosis in individual non-White 'patients' who presented with high-risk clinical profiles because they placed weight on knowledge that lung cancer risk factors and prevalence are lower in Black and South Asian than White populations.[39] However, there is no evidence that patients of different ethnicities exposed to the same risk factors with similar symptoms are at different risk of lung cancer so differential investigation by ethnicity is not clinically warranted. Another possible explanation is that investigation likelihood is influenced by GPs' ethnicity. In this study there were only seven GPs identified as Black, so it was not possible to examine this, but the mechanism by which observed ethnic variations in decision making occur remains an important question to address.

Strengths and limitations

Our novel approach, using vignettes in an interactive website that delivered real-time responses, obtained comprehensive information on decision making in over 99% of consultations and in a timeframe comparable to a typical consultation. The method simulated more components of the decision-making process in real time than has been achieved in previous studies.[40-42]

Of equal importance is the fact that we applied a randomised, factorial, experimental design, with exact balance on profile and risk, and approximate balance, with random allocation, to GPs, on socio-demographic factors. This allowed us to examine the effects of patients' socio-demographic and clinical characteristics on GPs' decision-making. We were not able to achieve total orthogonality in design of every patient characteristic, but the randomisation and approximate balance give some confidence in the general applicability of our results.

Despite the advances we achieved in simulating real consultations, the on-line vignettes were limited mainly due to the constraints of the natural language system. These constraints meant the website was unable to provide responses to all GPs' information requests. In the post-consultation survey 12 GP participants (5%) reported difficulty in obtaining information,

which caused some of them frustration, and a small number (n=4, 1.8%) observed it may have altered their decision-making behaviour. The process itself of typing in questions may also have prompted GP participants to consider their clinical reasoning more than they would in their routine clinical practice. Conversely, the opportunity to select from the extensive drop-down selections of examinations without facing any of the logistical constraints faced in a real consultation (e.g. time required to measure weight) may have led them to seek more information with less consideration than they would do in routine clinical practice. However, it is important to note that all approaches to simulating consultations have some drawbacks. For example, while other vignette studies have enabled physicians to 'ask' questions of the patient, this has required a researcher to type responses online as 'the patient', sometimes resulting in longer 'consultations' than real consultations.[40-42] Moreover, there are several reasons why these simulations still provide valuable insights into GPs' decision making. Firstly, our sensitivity analysis indicates that results were very close to the main analysis even after taking into account GPs' difficulties in obtaining responses from the application. Secondly, shortcomings in doctor-patient communication during the clinical encounter are well recognised, such that patients in real consultations do not volunteer all the information clinicians would need to make informed decisions.[18] Thirdly, it is the divergence from reality that makes simulated consultations useful for studying phenomena or circumstances not possible to observe or investigate in real life.[43] In this study, this divergence enabled the systematic manipulation of patient characteristics to examine their effects on GPs' decisions in isolation of the complex range of patient expectations and co-morbidities that might explain variations in decision making in real life. The divergence also meant GPs were not faced with the logistical and system/organisational constraints that affect referral decisions in practice. As a result, the findings provide insight into the cognitive processes underlying GPs' decision making when the variation in system and patient factors present in real life are removed.

There was some bias in the GP sample registering for the study in that GP participants' practices had higher cancer referrals than non-participating practices, so they may be more ready than GPs nationally to investigate symptoms suggestive of cancer. However, there was no evidence to suggest participating GPs would have greater or smaller variation in decision making than non-participants.

Another possible limitation is that the risk levels were based on positive predictive values from the CAPER symptom case-control dataset, which had wide and overlapping confidence intervals (as shown in supplementary data, S1). Therefore, the PPVs alone are not sufficient to conclude that clinical risk and therefore decision making should have varied by profile. However, even where the PPV point estimates are most disparate and confidence intervals overlap minimally, GPs investigated similar proportions of patients. In addition, the risk profiles had additional information other than PPV which should have guided decision making if GPs were acting in line with the latest available clinical guidance (e.g. symptom duration). Furthermore, our three broad categories align well with the 2015 NICE guidance. These equate to: risk below 1%, safety-netting; 1-3%, test in primary care if possible; over 3% refer for specialist testing.[29]

Conclusions and implications for research and practice

This study demonstrates that GPs were not more likely to initiate cancer investigations for 'patients' with higher risk symptoms. Furthermore, they do not investigate everyone with the same symptoms equally. It also indicates that insufficient data gathering could be responsible for diagnostic errors. It is not that GPs are doing a bad job: the average GP sees one patient with new lung cancer a year.[16] Distinguishing symptoms indicating possible cancer from self-limiting illness that GPs see daily, therefore is challenging. However, non-clinical variations in investigation could contribute to the socio-demographic inequalities in the timeliness of diagnosis and survival of lung cancer seen in the UK. It also marks a departure from the National Health Service commitment to promote equality through its services.[44] The findings also have wider implications for quality and safety in healthcare internationally. According to the Institute of Medicine, diagnostic errors contribute to approximately 10 percent of patient deaths, and sufficient data gathering is an essential part of reaching a timely diagnosis.[33]

It is therefore incumbent on health systems to consider strategies that can be implemented in practice such as clinician education,[33,45] decision support tools[25] and the assessment of equity in clinical practice.

REFERENCE LIST

1. Coleman MP, Forman D, Bryant H, et al. Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995-2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data. *Lancet* 2011;**377**:127-38.
2. Lyratzopoulos G, Neal RD, Barbiere JM, et al. Variation in number of general practitioner consultations before hospital referral for cancer: findings from the 2010 National Cancer Patient Experience Survey in England. *Lancet Oncol* 2012;**13**:353-65.
3. Neal RD, Din NU, Hamilton W, et al. Comparison of cancer diagnostic intervals before and after implementation of NICE guidelines: analysis of data from the UK General Practice Research Database. *Br J Cancer* 2014;**110**:584-92.
4. O'Dowd EL, McKeever TM, Baldwin DR, et al. What characteristics of primary care and patients are associated with early death in patients with lung cancer in the UK? *Thorax* 2015;**70**:161-68.
5. Walter FM, Rubin G, Bankhead C, et al. Symptoms and other factors associated with time to diagnosis and stage of lung cancer: a prospective cohort study. *Br J Cancer* 2015;**112**(s1):S6-S13.
6. Kostopoulou O, Delaney BC, Munro CW. Diagnostic difficulty and error in primary care—a systematic review. *Fam Pract* 2008;**25**:400-13.
7. Raine R, Wong W, Scholes S, et al. Social variations in access to hospital care for patients with colorectal, breast, and lung cancer between 1999 and 2006: retrospective analysis of hospital episode statistics. *BMJ* 2010;**340**:b5479.
8. Macleod U, Mitchell ED, Burgess C, et al. Risk factors for delayed presentation and referral of symptomatic cancer: evidence for common cancers. *Br J Cancer* 2009;**101 Suppl 2**:S92-s101.
9. Singh H, Sittig DF. Advancing the science of measurement of diagnostic errors in healthcare: the Safer Dx framework. *BMJ Qual Saf* 2015;**24**:103-10.
10. Hrisos S, Eccles MP, Francis JJ, et al. Are there valid proxy measures of clinical behaviour? A systematic review. *Implement Sci* 2009;**4**:37. doi: 10.1186/1748-5908-4-37.
11. Peabody JW, Luck J, Glassman P, et al. Comparison of vignettes, standardized patients, and chart abstraction: a prospective validation study of 3 methods for measuring quality. *JAMA* 2000; 283: 1715–22.
12. Kostopoulou O, Oudhoff J, Nath R, et al. Predictors of diagnostic accuracy and safe management in difficult diagnostic problems in family medicine. *Med Decis Making*. 2008;28:668-80. doi: 10.1177/0272989x08319958.
13. Zwaan L, Monteiro S, Sherbino J, et al. Is bias in the eye of the beholder? A vignette study to assess recognition of cognitive biases in clinical case workups. *BMJ Qual Saf* 2016. 10.1136/bmjqs-2015-005014
14. Blumenthal-Barby JS, Krieger H. Cognitive Biases and Heuristics in Medical Decision Making: A Critical Review Using a Systematic Search Strategy. *Med Decis Making*. 2015;**35**:539-57.
15. National Institute for Health and Care Excellence. Referral guidelines for suspected cancer: NICE guidelines CG27, 2005. Available: https://www.nice.org.uk/Guidance/CG27. Accessed 10 May 2016.
16. Hamilton W. The CAPER studies: five case-control studies aimed at identifying and quantifying the risk of cancer in symptomatic primary care patients. *Br J Cancer* 2009;**101 Suppl 2**:S80-6.
17. Adler MD, Trainor JL, Siddall VJ, et al. Development and evaluation of high-fidelity simulation case scenarios for pediatric resident education. *Ambul Pediatr* 2007;**7**:182-6.

18. Shim J, Brindle L, Simon M, et al. A systematic review of symptomatic diagnosis of lung cancer. *Fam Pract* 2014;**31**:137-48.
19. Doan S, Conway M, Phuong TM, et al. Natural Language Processing in Biomedicine: A Unified System Architecture Overview. In: Trent R, ed. Clinical Bioinformatics. New York, NY: Springer New York, 2014:275-94.
20. Spiegelhalter D, Abrams K, Myles J. *Bayesian Approaches to Clinical Trials and Health Care Evaluation*. Chichester, Wiley; 2004.
21. Stan Development Team. RStan: the R interface to Stan. Version 2.5; 2014. Available: http://mc-stan.org/rstan.html
22. Health and Social Care Information Centre. Workforce. 2014 Available: http://www.hscic.gov.uk/workforce. Accessed 10 May 2016.
23. National Cancer Intelligence Network. Practice Profiles. 2013 Available: http://www.ncin.org.uk/cancer_information_tools/profiles/gp_profiles Accessed 10 May 2016.
24. Hamilton W, Green T, Martins T, et al. Evaluation of risk assessment tools for suspected cancer in general practice: a cohort study. *Br J Gen Pract* 2013;**63**:e30-6.
25. Hamilton W, Peters TJ, Round A, et al. What are the clinical features of lung cancer before the diagnosis is made? A population based case-control study. *Thorax* 2005;**60**:1059-65.
26. Banks J, Hollinghurst S, Bigwood L, et al. Preferences for cancer investigation: a vignette-based study of primary-care attendees. *Lancet Oncol* 2014;**15**:232-40.
27. Department of Health. Improving Outcomes: A Strategy for Cancer. London: Department of Health, 2011. Available: www.gov.uk. Accessed 10 May 2016.
28. Hippisley-Cox J, Coupland C. Identifying patients with suspected lung cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract* 2011; 61: e715–23.
29. National Institute for Health and Care Excellence. Suspected cancer: recognition and referral. Secondary Suspected cancer: recognition and referral. 2015. Available: nice.org.uk/guidance/ng12. Accessed 10 May 2016.
30. Jensen H, Nissen A, Vedsted P. Quality deviations in cancer diagnosis: prevalence and time to diagnosis in general practice. *Br J Gen Pract* 2014;**64**:e92-e98.
31. Singh H, Daci K, Petersen LA, et al. Missed Opportunities to Initiate Endoscopic Evaluation for Colorectal Cancer Diagnosis. *Am J Gastroenterol* 2009;**104**:2543-54.
32. Lyratzopoulos G, Vedsted P, Singh H. Understanding missed opportunities for more timely diagnosis of cancer in symptomatic patients after presentation. *Br J Cancer* 2015;**112 Suppl**:S84-91.
33. Balogh E, Miller B, Ball JE. (eds) Improving diagnosis in health care / Committee on Diagnostic Error in Health Care. Washington, DC : The National Academies Press, 2015. DOI: 10.17226/2179442.
34. Zwaan L, Thijs A, Wagner C, Timmermans DR. Does inappropriate selectivity in information use relate to diagnostic errors and patient harm? The diagnosis of patients with dyspnea. *Soc Sci Med* 2013;**91**:32-8.
35. Kostopoulou O, Sirota M, Round T, et al. The Role of Physicians' First Impressions in the Diagnosis of Possible Cancers without Alarm Symptoms. *Med Decis Making*. 2016 doi:10.1177/0272989X16644563
36. McBride D, Hardoon S, Walters K, et al. Explaining variation in referral from primary to secondary care: cohort study. *BMJ* 2010;**341**:c6267. doi: 10.1136/bmj.c6267.
37. Tate AR, Nicholson A, Cassell JA. Are GPs under-investigating older patients presenting with symptoms of ovarian cancer? Observational study using General Practice Research Database. *Br J Cancer* 2010;**102**:947-51.

38. Scott SE, Walter FM, Webster A, et al. The model of pathways to treatment: conceptualization and integration with existing theory. *Br J Health Psychol* 2013;**18**:45-65.
39. Jack RH, Davies EA, Moller H. Lung cancer incidence and survival in different ethnic groups in South East England. *Br J Cancer* 2011;**105**(7):1049-53.
40. Kostopoulou O, Rosen A, Round T, et al. Early diagnostic suggestions improve accuracy of GPs: a randomised controlled trial using computer-simulated patients. *Br J Gen Pract* 2015;65:e49-54.
41. Hooper LM, Weinfurt KP, Cooper LA, et al. Virtual standardized patients: an interactive method to examine variation in depression care among primary care physicians. *Prim Health Care Res Dev.* 2008; 9: 257-268.
42. Harries C, Forrest D, Harvey N, et al. Which doctors are influenced by a patient's age? A multi-method study of angina treatment in general practice, cardiology and gerontology. *Qual Saf Health Care.* 2007; 16: 23-27.
43. Hughes R. Considering the Vignette Technique and its Application to a Study of Drug Injecting and HIV Risk and Safer Behaviour. *Sociol Health Ill* 1998;**20**:381-400.
44. Department of Health. Equity and Excellence: Liberating the NHS. 2010. Available: www.gov.uk. Accessed 10 June 2016.
45. Reilly JB, Ogdie AR, Von Feldt JM, et al. Teaching about how doctors think: a longitudinal curriculum in cognitive bias and diagnostic error for residents. *BMJ Qual Saf* 2013;**22**:1044-50.

FIGURES AND LEGENDS

Figure 1. Annotated screen grab from the study website showing an example 'patient', Jack Jones' (Profile 1) thumbnail sketch: White British man, aged ~60 years

LEGEND: 'Jack Jones' (Profile 1) thumbnail sketch: White British man, aged ~60 years, works on security in a block of offices. Non-smoker, has a history of diabetes (*available through drop-down menu and shown in video if asked)*

- Social characteristics: Socioeconomic circumstances - not stated directly but GPs indicated in the accent and dress of 'patients' and occupation provided on a drop-down menu. Ethnicity - listed in a drop-down menu and reflected in vignette patients' accent and dress, with racial characteristics apparent in the facial features of the 'patient'. All actors spoke fluent English to avoid linguistic barriers.
- Presenting symptom *(shown to all)*: Breathless. Never felt like this before and he is not sure what's going on. It's interfering with his life (e.g. had to get the bus into work rather than walking) and so wife suggested he come and check it out.
- Second symptom (*available if asked*): Fatigue. Presumed this is because of breathlessness, but it is more severe than normal. Not sure why: work is the same as normal, things are no different at home, and he doesn't feel stressed.
- Further information on core profile characteristics (*available if asked)*: Notice it particularly when active (e.g. struggle playing with the grandchildren). Also notice it when lying down in bed, and has had to start using one of his wife's pillows. It's been happening for 1-2 weeks (e.g. trains young boxers at the local gym but hasn't been able to make boxing training for the last week and a half because of it).
- Examples of additional profile information (*available through drop-down menus*): medications, heart rate, blood pressure, weight, height, joints (all OK).

SUPPORTING INFORMATION FILES

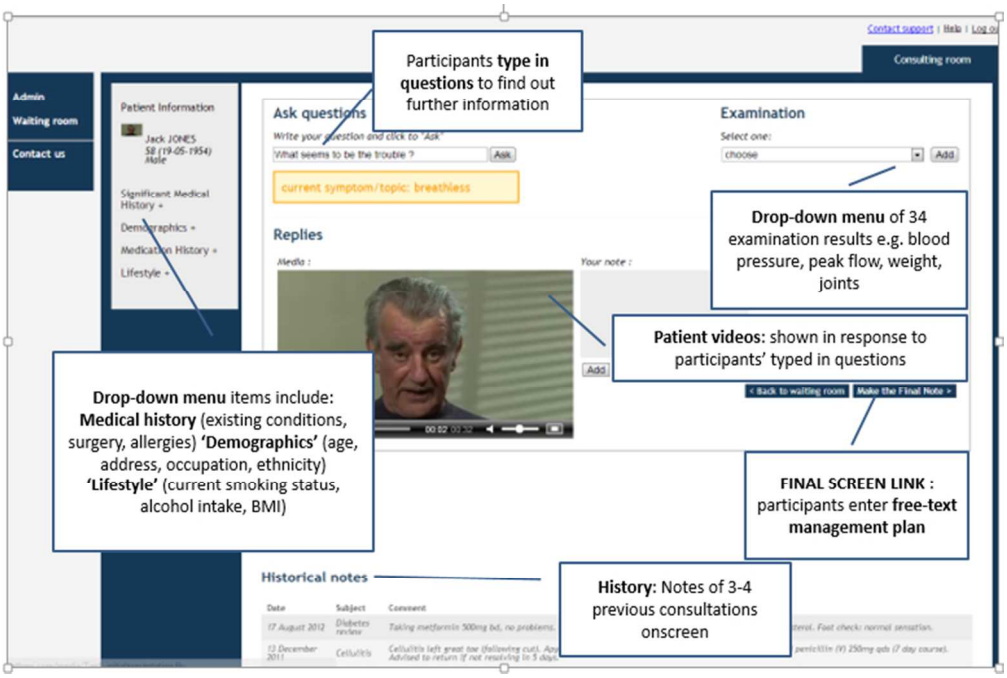
S1. Methods: Application development

S2. Results: Additional details

A. Participant characteristics

B. Comparison of sample with national GP population
C. Characteristics of GPs' simulated consultations and experience of the application
D. Sensitivity analysis including error messages received as a co-variable in both models

S3. **Results: GP comments on the study design**



Annotated screen grab from the study website showing an example 'patient', Jack Jones' (Profile 1) thumbnail sketch: White British man, aged ~60 years

65x43mm (300 x 300 DPI)

Supplementary file S1

METHODS

Steps in developing the virtual patient application

Step 1. Case concept: developing the vignette design and content

The factorial experimental design, informed by reviews of the literature on non-clinical factors affecting GP decision making and lung cancer diagnosis and survival, covered four experimental factors (**Table 1)** known to be associated with variation in lung cancer survival, but whose effect on inequalities in GPs' rates of referral for investigation or to secondary care is uncertain:[1]

- **Ethnicity**: three variations (White, Black Caribbean, South Asian)
- **Gender**: two variations (male, female)
- **Socioeconomic circumstances**: two variations (advantaged or disadvantaged)
- **Clinical risk of lung cancer**: three variations (low, medium and high risk), with two profiles for each level of risk. Age was not included as a separate experimental factor but was instead incorporated into profiles because older age increases the risk of cancer associated with most symptom combinations.[2]

We constructed six clinical profiles, two for each risk level using different combinations of symptoms, age, and smoking status.(**Table 2)** The clinical profiles and risk levels were based on positive predictive values (PPV) of lung cancer from the CAPER case-control dataset and the latest available NICE guidance on investigation of suspected cancer. [2 3] This 2005 guidance recommended investigation for symptoms present for >3 weeks - which equates to PPVs >1.2%, though the guidance did not specifically use PPV thresholds. The low-risk vignettes (Profiles 1 and 2) therefore reflected a PPV<1.2% with symptom duration of 1-2 weeks, such that investigation would not be indicated (**Table 2**). Investigation would be clearly indicated for 'high-risk vignettes (Profiles 5 and 6), which reflected a PPV>3% with symptom duration more than 5 weeks. In the medium-risk vignettes (Profiles 3 and 4, PPV=1.7-2.5%), investigation would be consistent with guidance but information on symptom duration was kept intentionally vague so "safety-netting" (i.e. a back-up plan if symptoms persist or escalate) without active investigation could also be appropriate. The first symptom was volunteered by the 'patient,' the second only elicited if GPs specifically asked. In Profile 1 (low risk), symptoms and co-morbidities unrelated to lung cancer, to deflect GPs from the primary purpose of the study.

For each clinical profile a comprehensive set of additional information was developed to include:

- Medical records for each of the 'patients'; similar to what GPs would find in their own clinical system. These incorporated information on socio-demographic and lifestyle

characteristics, past medical history and medication, and a recent consultation history. Consistent with what would be expected for patients of their age, many 'patients' had co-morbidities; but these were selected so in most cases they did not alter the patient's likelihood of lung cancer.

- Results of examinations and tests that GPs might perform: including tests unrelated to the risk profile symptoms to avoid priming GPs' behaviour. In most cases results were the same for all 'patients' with that profile, although some varied according to 'patient' gender. The respiratory and cardiovascular examinations were unremarkable for all six profiles to ensure we were studying GPs' responses to the presence/absence of symptoms, rather than to positive examination findings.

Step 2. Review and Revision by Content Experts

To maximise the clinical authenticity of the cases, GPs specializing in cancer diagnosis and non-academic GPs reviewed the proposed vignettes, which were then revised following their comments. The website content and functionality were also informed by patient representatives' comments. This led to the inclusion of non-smokers because of the risk of diagnostic delay if GPs are less likely to suspect cancer in this group. It also directly informed the types of responses 'patients' provided, where patient representatives corroborated previous research that patients may well not disclose certain symptoms with their doctors without being directly asked about them.

Step 3. Outline and Flow Development: A typical consultation in the study

We developed the outline flow of a typical consultation on the application, the duration of which would be determined by what the GP sought to find out:

- GP enters "waiting room" and clicks on "Patient" link
- GP selects question "What seems to be the trouble?"
- Video plays where the 'patient' volunteers their first symptom; the view is a head shot of 'patient' in GP consulting room.
- GPs can find out additional information through:
    - asking the 'patient' questions (e.g. on the nature of a symptom, presence of other symptoms). 'Patient' videos then play giving the GP requisite information. If the system was unable to provide a meaningful response, users receive an error message.
    - consulting medical records (e.g. on previous consultations, medications),
    - performing examinations (e.g. blood pressure, with findings provided as text).
- GP selects "Make the final note" where they enter ideas about diagnosis (main, possible, possible but unlikely) and their management plan.
- The GP completes six such consultations over ~3 weeks.

Table 1: Each of the 36 vignettes combinations representing the four experimental factors: gender, socioeconomic circumstances, ethnicity, and risk level (across two clinical profiles)

	Low risk (PPV <1.2%)		Medium risk (PPV = 1.7-2.5%)		High risk (PPV ≥ 3%)	
	Clinical Profile 1 PPV=0.4 [0.1-3.1]	Clinical Profile 2 PPV=0.4 [0.1-3.1]	Clinical Profile 3 PPV=0.4 [0.1-3.1]	Clinical Profile 4 PPV=0.4 [0.1-3.1]	Clinical Profile 5 PPV=0.4 [0.1-3.1]	Clinical Profile 6 PPV=0.4 [0.1-3.1]
Disadvantaged	1. South Asian	2. South Asian	3. South Asian	4. South Asian	5. South Asian	6. South Asian
	7. Black	8. Black	9. Black	10. Black	11. Black	12. Black
	13. White	14. White	15. White	16. White	17. White	18. White
Advantaged	19. South Asian	20. South Asian	21. South Asian	22. South Asian	23. South Asian	24. South Asian
	25. Black	26. Black	27. Black	28. Black	29. Black	30. Black
	31. White	32. White	33. White	34. White	35. White	36. White

Key:	Male	Female

Table 2. Components of the six different clinical profiles by risk level

Clinical Profile	Information volunteered by 'patient' or available onscreen			Information only available if participant GPs asked		Positive Predictive Value (PPV) of lung cancer	Other relevant information
	Age range	Smoking status	Symptom 1	Symptom 2	Duration		
Low risk: Expected action = no active investigation (safety netting appropriate)							
1	Younger (Late fifties)	Non smoker	Breathlessness	Fatigue	1-2 weeks	0.40%	Patient has swollen ankles, possibly due to heart failure
2	Younger (Late fifties)	Smoker	Chest pain	Cough	1-2 weeks	1.10%	
Medium risk: Expected action = either investigation (e.g. order chest x-ray) or safety netting							
3	Older (Late seventies)	Smoker	Chest pain	Cough	Uncertain (approx 3 weeks)	1.70%	
4	Older (Late seventies)	Non-smoker	Cough	Appetite loss	Uncertain (approx 3 weeks)	2.50%	
High risk: Expected action = lung cancer investigation							
5	Younger (Late fifties)	Smoker	Breathlessness	Fatigue	>5 weeks	3-4%	Chronic obstructive pulmonary disease (COPD) present
6	Older (Late seventies)	Smoker	Chest pain	Weight loss	>5 weeks	14%	

Step 4. Translation of vignette content into simulation platform: the virtual patient application website

The translation of content into the virtual patient application website required filming actors portraying patients, creating and populating the website with that content. Twelve actors with medical role-playing experience were each filmed playing three 'patients' in a studio resembling a GP's consulting room. Actors were selected to fulfil the 'patient' template of the factorial design, i.e. every combination of three ethnicities and male/female across the two age groups (58/59 year olds, and 78/79 years). Each actor received a detailed brief for three 'patients'. This contained profile information (e.g. symptom presentation and features) plus details relating to the specific character (e.g. occupation). Actors represented the socio-economic circumstances of their 'patient' profiles through appearance (e.g. clothing, hairstyle, makeup), accent and information about their occupation/lifestyle. In each case actors started with an introduction to their presenting symptom - how one might answer a GP's initial question, "What seems to be the trouble?" – and continued with responses to additional questions about specific features of the presenting symptom, additional symptoms and their features, and other relevant subjects (e.g. smoking status). Actors were asked to describe these symptoms in their own words but had example scripts provided by patient representatives of how real patients might describe their experiences and sensations. To ensure consistency in content across all the vignettes, the researchers used checklists to ensure the actors had mentioned all the details relevant to their profiles.

The website architecture and application software was produced by Athenaeum Educational Technologies. It involved the development of a bespoke system using natural language processing principles to recognise GPs' free-text questions and play a video clip in response (see Doan et al 2014 for an explanation of the principles).[4] This system was underpinned by databases on symptoms or risk factors and the features those symptoms (e.g. what exacerbates or relieves the symptom or how long it has been present). Each database was populated with a set of key words (including common typographical errors) which GPs might use to ask about the existence and features of these symptoms, see Table 3 for an example. The key words were initially developed by the research team in consultation with content experts and subsequently extended to enable the system to respond to the language and content of questions asked by GPs in pilots. Finally, the website was populated for each vignette with:

- Over 300 videos of the 'patient' actors describing symptoms and main risk factor responses to provide answers to GPs' typed in questions. This included a generic "No" or "Don't understand" where there was no clinically relevant information available.
- Text (available as drop-down menu) for all other aspects.

Table 3. Database example: Key words for Dyspnoea (symptom) and onset (feature)

symptom or risk factor	shortness breathless breathlessness breathe dyspnoea puff short of breath lost breath lose breath catch breath breatlessness breatless breathing difficulty breathing trouble breathing out of breath	Onset	what brings exacerbates what triggers makes it happen start to happen causes exacerbate aggravate aggravates agrivate aggrivate agrivates aggrivates especially bad aggrevate aggrevates makes it worse exacerbation pleuritic plueritic deep breath taking a breath take a breath breathing in breathe in breath in

Step 5. Pilot testing

Three pilot stages were conducted to identify changes needed to content, functionality and design. In stage 1, researchers were present whilst three GPs tried up to four online consultations to identify any problems in using the application and where additional vignette content was needed. In stage 2, GPs (n=7) conducted up to four online consultations remotely. After their pilot, they participated in a telephone interview with a researcher to provide feedback on the intuitiveness of the application, credibility of the vignettes, the consultation process and the extent to which they were able to use similar reasoning as in their day-to-day practice. In stages 1 and 2, researchers (JMc, RS, JS) reviewed participants' log forms to identify where GPs' questions led to an error message or an inappropriate video response. Revisions to the website databases and functionality were revised in response. In stage 3, researchers not connected with the study (n=10) conducted up to four consultations

to check whether errors from pilots 1 and 2 had been corrected. See S1 for details of revisions made as a result of the pilots.

We made changes following pilots to the appearance, content, and functionality of the application:

- **Appearance**: We altered the design after the first pilots to ensure GPs could see all the opportunities to find out information on onscreen without scrolling down and changed the colour scheme in response to pilot feedback.

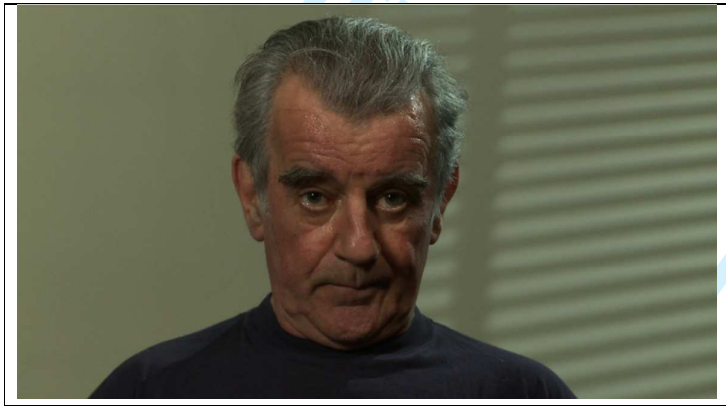
  Content: In response to GP feedback in the early pilots, we filmed longer 'patient' clips describing symptoms (from 15-30 secs to 45-60 secs) with less relevant clinical information (from describing the nature and frequency of symptoms to just reporting presence of symptom and instead recounting effects on daily life). In addition, we added more content for each profile was developed and filmed to provide answers to a wider range of questions.

  Function: Using the log file data from the plots, the symptom and features databases were extended and refined to enable the website to provide more meaningful answers to GPs' questions.

  There were limitations in natural language function that could not be further overcome. For example the application required GPs to repeat the name of the symptom they were asking about in all their questions (e.g. 'how long have you had chest pain' or 'what makes the breathlessness worse') which does not realistically mimic spoken conversation. We used data from the log files on where these caused GPs to get error responses in the pilots to inform development of a help video and PDF that GPs could access whilst using the application. We also provided GPs with feedback after their first 'consultation' to reduce the likelihood that they missed key information in future 'consultations' because of repeated error.

Profile 1: complete vignette

The description below illustrates the information available for each profile and how GP participants could access it. It is shown for the first vignette profile, the 'deflecting' vignette, where the risk of lung cancer is lowest and the most probable diagnosis is heart failure

	Thumbnail: 'Jack Jones' White British man, aged ~60 years, works on security in a block of offices. Non-smoker, has diabetes and a history of depression,

Data item	Accessed by	Information	Format
Presenting symptom	Video – Displayed when participant clicked on default question on screen	Breathless	Patient account: *Never felt like this before and he is not sure what's going on. It's interfering with his life (e.g. had to get the bus into work rather than walking) and so wife suggested he come and check it out.*
Second symptom	Video – Displayed if participant used text box to ask a direct question about presence of symptom. *Synonyms recognised included: tiredness, tired Energy, lethargic, lethargy,*	Fatigue	Patient account: *Presumed this is because of breathlessness, but it is more severe than normal. Not sure why: work is the same as normal, things are no different at home, and he doesn't feel stressed.*

	drained, exhaustion *Exhausted, fatigue, fatigued, sluggish, knackered* *pooped*		
Further information on symptom characteristics	Video – Displayed if participant used text box to ask a direct question about characteristics of the symptom. *For breathlessness, questions that could be addressed included: How long have you been breathless?* *What makes it better?* *What makes it worse? How far can you walk?* *Is it worse on exercise?* *Is it worse when you lie down?* *Does it stop your normal activities?* *Can you carry things?* *Have you ever had this before?* *Do you have chest pain?* *Do you have swollen ankles?* *Have you had calf swelling?* *Do you have asthma?* *Do you have COPD?* *Are you a smoker?* *Do you have heavy periods?)*	• Duration • Onset • Offset • Frequency • Effect of: exercise, lying down • life changes • Diet, bowel • Position (of pain) • Illness ideas • Family history • Medication • Related symptoms	Patient account: *Notice it particularly when active (e.g. struggle playing with the grandchildren). Also notice it when lying down in bed, and has had to start using one of his wife's pillows. It's been happening for 1-2 weeks (e.g. trains young boxers at the local gym but hasn't been able to make boxing training for the last week and a half because of it).*

Information available from drop-down menus

Patient information		Comments
Name	Jack Jones	Consistent with ethnicity and socioeconomic circumstances
Date of birth	19.05.1954	
Gender	Male	Also evident from patient video
Address (first line)	xxx	Consistent with socio-economic circumstance as far as possible…
Ethnicity	White	Also evident from patient video (dress, accent)
Occupation	Security guard	Also available as patient video (in response to questions about occupation, job etc)
Lifestyle factors		
Smoking status	Never smoked	Also available as video if patient asked through text box
Units of alcohol per week	25 units	Consistent with socioeconomic and ethnicity profile and set so as not to raise suspicion that alcohol misuse caused symptoms.
BMI	xx kg/m²	See weight
Family history	None recorded	

Systems Examinations	Information
Abdomen (including rectal)	Soft and non tender. No abnormalities detected.
Breast	Normal.
Cardiovascular system (note to us, include heart rate/rhythm)	Heart rate 72 beats/minute. Regular rhythm. Normal hearts sounds. No sacral or peripheral oedema.
ENT examination	No abnormality detected.
Eye examination (including fundoscopy)	No abnormality detected.
Foot examination	Pulses palpable. Sensation normal.
Genitalia examination	No abnormality detected.
Heart rate/rhythm	Heart rate 72 beats/minute. Regular rhythm.
Nail examination	All nails appear normal.

Neurological examination, central (including cranial nerves)	No abnormality detected.
Neurological examination, peripheral	No abnormality detected.
Peripheral pulses	All pulses palpable. No abnormality detected.
Respiratory rate	14 breaths/minute.
Respiratory system	Rate: 14 breaths/minute. No peripheral or central cyanosis. Good chest movement. Chest clear.
Joint examination, cervical spine	Good range of pain-free movement.
Joint examination, shoulder	Both joints normal in appearance and movement.
Joint examination, elbow	Both joints normal in appearance and movement.
Joint examination, wrist	Both joints normal in appearance and movement.
Joint examination, hand	Joints normal in appearance and movement.
Joint examination, thoraco-lumbar spine	Normal gait. Good range of pain-free movement.
Joint examination, hip	Both joints normal in appearance and movement.
Joint examination, knee	Both joints normal in appearance and movement.
Joint examination, ankle	Both joints normal in appearance and movement.
Joint examination, foot	Joints normal in appearance and movement.
Bedside tests	**Information**
Blood glucose	6.7 mmol/L
Blood pressure	140/80 mmHg
Cultures	Sputum sample provided and sent to laboratory
Height	180cm (men); 163cm (women)
Peak flow	575l/min (men); 390l/min (women)
Swabs	Swabs taken and sent to laboratory
Temperature	36.5^{0}C
Urinalysis	Urinalysis normal
Weight	What seems reasonable for actor/actress

Significant medical history	**Information**
Co-morbidities & date of diagnosis	Diabetes mellitus 24.11.09
	Depression 05.01.11
	Allergies None recorded
Current medication	**Information**
Drug name, dose, instructions for use	Metformin 500mg bd
	Fluoxetine 20mg od
Recent appointment history	
	01.08.12 Diabetes Review Taking metformin 500mg bd, no problems. BP: 140/90 (on ramipril, amlodipine). HbA1c: 7.1. Normal FBC, renal function, cholesterol. Foot check: normal sensation.
	25.10.11. Cellulitis. Cellulitis L great toe (following cut). Apyrexial, does not appear unwell. Rx: flucloxacillin 250mg and penicillin (V) 250mg qds (7 day course). Advised to return if not resolving in 5 days. 09.08.11 Diabetes Review Taking metformin 500mg bd, no problems. HbA1c: 7.5. Foot check: normal. Discussed dietary compliance.

References

1. Macleod U, Mitchell ED, Burgess C, et al. Risk factors for delayed presentation and referral of symptomatic cancer: evidence for common cancers. Br J Cancer 2009;**101 Suppl 2**:S92-s101.
2. Hamilton W. The CAPER studies: five case-control studies aimed at identifying and quantifying the risk of cancer in symptomatic primary care patients. Br J Cancer 2009;**101 Suppl 2**:S80-6.
3. NICE. Referral guidelines for suspected cancer: NICE guidelines CG27, 2005.
4. Doan S, Conway M, Phuong TM, et al. Natural Language Processing in Biomedicine: A Unified System Architecture Overview. In: Trent R, ed. Clinical Bioinformatics. New York, NY: Springer New York, 2014:275-94.

S2 Additional results

A. GP Participant characteristics

	n	%
Gender		
Female	101	44.69
Male	125	55.31
Age range		
25-34 years	54	23.89
35-44 years	77	34.07
45-54 years	70	30.97
55-64 years	23	10.18
65 years or over	2	0.88
Years since qualifying		
0 to 2 years ago	28	12.39
2 to 5 years ago	45	19.91
5 to 10 years ago	40	17.70
10 to 20 years ago	55	24.34
20 or more years ago	58	25.66
Ethnicity		
White (British, Irish, Other)	132	58.41
South Asian (Indian, Bangladeshi, Pakistani or Asian mixed)	67	29.65
Black (Afrian or Caribbean)	7	3.1
Other or missing (eg Chinese)	20	8.85
Region		
London	81	35.84
East of England	76	33.63
North West	29	12.83
West Midlands	22	9.73
Surrey and Sussex	9	3.98
Locum GP	6	2.65
Total	**226**	

B. Participant vs national general practice characteristics

Practice population and cancer referral characteristics	All median (IQR)	Nonparticipant practices median (IQR)	Participant practices median (IQR)	p
Age standardised cancer referral ratio	83 (37-156)	83 (36-155)	118.5 (55.25-192)	<0.001
Proportion of practice pop'n over 65y	15.4 (11.3-18.6)	15.4 (11.4-18.6)	13.1 (8.7-18.2)	0.005
n	**8365**	**8145**	**220**[(b)]	

[(b)]N is not 226 because 6 participants were locums so could not be assigned to a practice

GP characteristics	All GPs		Participants	
	n	%	n	%
Age				
25-34 years	4389	12.35	53	23.66
35-44 years	10920	30.74	78	34.82
45-54 years	12205	34.35	69	30.80
55-64 years + or unknown	8013	22.55	24	10.71
Gender				
Female	16,723	47.07	102	45.54
Male	18,804	52.93	124	55.36
Total	**35,527**		**226**	

C. Characteristics of GPs' simulated consultations and experience of the application

	Consultation 1	Consultations 2-6	All Consultations
Median duration [IQR]	13 [10-21] minutes	11 [8-14] minutes	
Mean number of information requests [SD]	46.1 [19.2]	47.7 [13.9]	47.4 [13.3]
Median % information requests resulting in error responses [Full range]	3.6% [0 - 52.9]	4.6% [4.0- 22.2]	4.8% [0 - 21.1]
n (consultations)	**227**	**1121**	**1348**

Note: the error rate for all consultations is higher than both consultations 2-6 and consultation 1 because in consultation 1 there were more cases where errors were low or zero but numbers of questions asked were also low than in later consultations.

D. Multilevel logistic regression of cancer investigation – adjusted for error rate experienced by GPs

	Adjusted odds ratio	CI (*= p<0.05)
Patient' characteristic		
Profile		
Profile 1 (PPV=0.8%)	0.30	
Profile 2 (PPV=1.1%)	0.82	[0.52; 1.3]
Profile 3 (PPV=1.7%)	1.00	
Profile 4 (PPV=2.5%)	0.20	[0.13; 0.31]*
Profile 5 (PPV=3-4%)	0.79	[0.49; 1.29]
Profile 6 (PPV~14%)	0.31	[0.2; 0.46]*
Ethnicity		
White	1.00	
South Asian	0.82	[0.62; 1.08]
Black	0.74	[0.57; 0.99]
Error rate[(d)]		
>10%	0.77	[0.5; 1.18]
≤10% queries resulted in error	1.00	

[(d)]% GP participant queries resulting in error message

S3. Supporting information_Results: GP comments on the study design

Overall, 24 GPs (11.5%) who completed the GP decision making study volunteered comments on the 'virtual patient application' study method or design using the free text question in a survey after the study or directly to the study team.

We share the themes emerging from these volunteered comments to provide insight into GPs' experiences on participating in the study and perspectives on the study design as a tool for examining decision making. However, this was not a survey of all participant GPs' views and experiences, therefore we cannot conclude that the views are representative of all participants or that others would not have expressed similar views if they have been asked directly about the study design.

Difficulties extracting information expected to be there

12 GP participants (5.3%) commented it was difficult to use the study tool to extract the information they would have wanted to receive. Most of these GPs commented that they experienced difficulty working out how to phrase questions to the 'patient' in order to play videos answering the question they wanted, which may have required changing their normal open questioning style:

- *'I did not find the online consultations easy to follow. I wanted to ask questions but did not know how to phrase them.' [GP 77]*
- *'I found the study quite frustrating because I was often unable to ask the questions I would normally ask and so did not obtain as good a history as usual and so felt I was making decisions with only half the information I normally have available.' [GP 15]*
- *'The vignettes are out of keeping with my style of open questions, so I found this difficult to explore symptoms.' [GP 65]*

Consultation behaviour diverged from 'real-life' due to application difficulties

Four GPs (1.8%) observed their consultation behaviour in this study diverged from their normal behaviour in ways that might have affected their decision making, perhaps leading them to under or over investigate 'patients':

- *'The frustration surrounding the uncertainty of the answers definitely lowered my threshold to refer and review again.' [GP 170]*
- *'Getting lots of no's or I don't know mean I felt a bit frustrated and gave up on the consultation.' [GP 77]*
- *'Wasting time trying to get the relevant history when the computer could not respond de-motivated me to engage or care if I performed well.' [GP 112]*
- *'I felt I may have over-investigated as unable to obtain answers to [certain] questions.' [GP 107]*

Differences between online simulations to real life

Some GPs also observed that (even if they were able to receive the information they would have sought from a real-life patient), simulated consultations online were different in important ways to real life consultations. In real life, GPs have the opportunity to pick up visual cues from seeing patients walking into the room, they are influenced by other contextual factors and they always have the opportunity to see patients again:

- *'I think a lot of what we learn comes from visual cues or other things within the consultation - e.g. how breathless they are walking into the room.' [GP 77]*
- *'It also makes it different when you actually see someone face to face.' [GP 187]*
- *'Each patient is an individual - your scenarios were difficult to put in a realistic context to make a valid assessment of what I personally would do in real life.' [GP 101]*
- *'There is a lot of contextual material in the decision to refer for tests and further opinions. Much of that could not be captured in these vignettes.' [GP 67]*
- *'History taking in practice is easier than the vignettes and often an option would be seeing [the patient] again.' [GP 139]*