What are we missing by ignoring text records in the Clinical Practice Research Datalink?

Using three symptoms of cancer as examples to estimate the extent of data in text format that is hidden to research

Submitted by Sarah Jane Price to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Medical Studies in January 2016

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature: ……..

Abstract

Electronic medical record databases (e.g. the Clinical Practice Research Datalink, CPRD) are increasingly used in epidemiological research. The CPRD has two formats of data: coded, which is the sole format used in almost all research; and free-text (or 'hidden'), which may contain much clinical information but is generally unavailable to researchers. This thesis examines the ramifications of omitting free-text records from research.

Cases with bladder (n=4,915) or pancreatic (n=3,635) cancer were matched to controls (n=21,718, bladder; n=16,459, pancreas) on age, sex and GP practice. Coded and text-only records of attendance for haematuria, jaundice and abdominal pain in the year before cancer diagnosis were identified. The number of patients whose entire attendance record for a symptom/sign existed solely in the text was quantified. Associations between recording method (coded or text-only) and case/control status were estimated (χ^2 test). For each symptom/sign, the positive predictive value (PPV, Bayes' Theorem) and odds ratio (OR, conditional logistic regression) for cancer were estimated before and after supplementation with text-only records.

Text-only recording was considerable, with 7,951/20,958 (37%) of symptom records being in that format. For individual patients, text-only recording was more likely in controls (140/336=42%) than cases (556/3,147=18%) for visible haematuria in bladder cancer (χ^2 test, $p<0.001$), and for jaundice (21/31=67% vs 463/1,565=30%, $p<0.0001$) and abdominal pain (323/1,126=29% vs

397/1,789=22%, p<0.001) in pancreatic cancer. Adding text records reduced PPVs of visible haematuria for bladder cancer from 4.0% (95% CI: 3.5–4.6%) to 2.9% (2.6–3.2%) and of jaundice for pancreatic cancer from 12.8% (7.3–21.6%) to 6.3% (4.5–8.7%).

Coded records suggested that non-visible haematuria occurred in 127/4,915 (2.6%) cases, a figure below that generally used for study. Supplementation with text-only records increased this to 312/4,915 (6.4%), permitting the first estimation of its OR (28.0, 95% CI: 20.7–37.9, p<0.0001) and PPV (1.60%, 1.22–2.10%, p<0.0001) for bladder cancer.

The results suggest that GPs make strong clinical judgements about the probable significance of symptoms – preferentially coding clinical features they consider significant to a diagnosis, while using text to record those that they think are not.

List of Contents

List of Tables

List of Figures

1 Introduction

The Clinical Practice Research Datalink (CPRD) is a UK-based research service providing anonymised copies of primary care records. It is considered the gold standard of primary care databases, partly because of its large size and generalisability to the UK population. Though created for pharmacological research, the database is now used extensively in epidemiological studies.[1,2]

One concern for researchers using the CPRD arises from the way clinical events are recorded electronically. Most CPRD practices use ViSion (ViSion INPS, London, UK), in which GPs must choose a Read code[a] first to begin a record, after which a comments box opens. Here, GPs can type freely and are not limited to elaborating on the code.[3] From clinical and medico-legal standpoints, codes and text are equally accessible. The same cannot be said for research – while codes are fully and routinely available to researchers, text records are not. Furthermore, a moratorium on collection of CPRD text data was introduced in 2013. It remains possible to access text recorded before then, although this is complex, expensive, limited and, therefore, rarely done.[b] Thus, the clinician's recording style may generate bias – researchers being oblivious

[a] Read codes are a coded thesaurus of clinical terms that have been in common use in the NHS since 1985.

[b] Note: on 31 March 2016, shortly after my viva, the CPRD stopped providing historical free text (that collected before 2013).

to anything that is only recorded as an inaccessible comment (henceforth called ‘hidden text’).

The objective of this thesis is to investigate the evidence for recording style bias in the CPRD, using, as an example, recently developed research methods underpinning risk assessment tools for GPs to assess cancer risk in symptomatic patients.[4,5] The implications of the results for another cancer prediction tool, QCancer, are also discussed.[6,7]

The objectives are achieved by re-creating datasets from case–control studies characterising the presentation of cancer in primary care, and supplementing them with text records for known symptoms of cancer. Three symptoms and two cancers were chosen to exemplify recording style bias. Haematuria and jaundice are high-risk 'alarm' symptoms of bladder and pancreatic cancer, respectively, and abdominal pain is a 'low-risk but not no-risk' feature common to both cancer sites.[4,5]

To meet the objectives, my study addresses five specific research questions:

1. How much symptom information is documented in electronic medical records using text rather than a code?

2. Are studies of coded data vulnerable to bias arising from the differential use of text and codes between comparison groups?

3. Does recording style vary with type of symptom?

4. Does the recording style vary with the clinical context of a symptom's presentation?

5. Do text data provide additional value to coded data?

To answer these questions, I quantify recording style bias and its effect on measures of association between the symptoms and cancer and on risk estimates for cancer (such as the positive predictive value).

The thesis begins with an overview of the CPRD, including its history, data formats, strengths and limitations, and its use in research. Chapter 3 reviews the literature on the types of information recorded in the CPRD as hidden text. Chapter 4 describes the background to this work – including the Discovery Programme, within which this project is set. It also gives brief overviews of bladder and pancreatic cancer, and of the three symptoms – haematuria, jaundice and abdominal pain. Chapter 5 summarises the main background to the project and identifies the research questions addressed by my thesis.

The methods are described in Chapter 6 and the results in Chapters 7 to 12. Chapter 13 is the discussion, and conclusion forms Chapter 14. The thesis ends with a number of appendices containing reference materials. These materials include my publication in *The British Journal of General Practice*, conveying my clinical findings about the risk of bladder cancer in patients with non-visible haematuria.[8] These results were used by The National Institute of Health and Care Excellence in their revision of the guidelines for recognition of cancer in primary care and referral for investigation.[9]

2 The Clinical Practice Research Datalink

2.1 History

The Clinical Practice Research Datalink (CPRD) (www.cprd.com) is a UK-based research service that maintains a database of anonymised, electronic copies of longitudinal patient records in primary care. Set up for pharmacological research, the database was originally called the Value Added Medical Products (VAMP) research databank, but was renamed the General Practice Research Database (GPRD) in 1993 when VAMP was bought out by Reuters Health Information. A year later the GPRD was donated to the Department of Health.[10] The rebrand to CRPD occurred in 2012 when the NHS National Institute for Health Research and the Medicines and Healthcare products Regulatory Agency assumed a joint funding role. For simplicity's sake, I shall use the term CPRD consistently throughout this thesis, regardless of the name by which the database was known at the time of the study under discussion.

2.2 Data recording

Since the 1990s, clinical data in the CPRD have been recorded within contributing GP practices during patient consultations using ViSion software. ViSion contains Structured Data Areas, in which Read codes are used to describe clinical information, 'to aid rapid and more complete recording of presenting symptom(s), [and] significant physical findings (positive and negative)…'.[3] Clinical data that are not entered as Read codes are captured as

text, primarily in the notes written by GPs to supplement or clarify Read codes (Figure 2.1), and in communications to and from other healthcare providers (for example, hospital letters). When writing the former, GPs are not limited to elaborating solely upon the clinical event that prompted their particular choice of code; hence, the text is commonly described as 'free'. Use of codes rather than text fields is encouraged, not only to fulfil the CPRD's commitment to provide high-quality data for research, but also because it enables automation of clinical audit within a practice.[3]

Figure 2.1 Screenshot showing the comments box that becomes available (red arrow) once a Read code has been selected in ViSion

2.2.1 Use of codes versus text

It is worth mentioning at this point that little is known about the risks and benefits of coding the patient history in electronic health records, or indeed about what motivates GPs in their choice of recording style, i.e. codes or text. Since the 1960s, advances in medical bioinformatics have resulted in a number of systems that enable structured (i.e. coded) entry of clinical information, but the data on their dissemination and use are scarce. Hamilton *et al* (2003)

reported on the quality of individual consultation recording in paper-only, computer-only and in hybrid systems (i.e. a combination of both). The quality of consultation recording was highest for paper-only systems, particularly regarding symptom recording.[11] A qualitative study published 8 years later suggested that structured data entry remained less flexible and more time-consuming than traditional text entry. [12] Advances in technology that address these objections will broaden the use of coding, as will improvements in the IT skills of doctors as younger generations progress through the system.

Furthermore, a systematic review of secondary care records concluded that structuring the medical history increases the amount of clinical information gathered and makes it more amenable to being coded. However, there was only poor evidence that coding is associated with improved clinical decision-making and no evidence that it is accompanied by an improvement in patient outcomes.[13] The review noted that the lack of evidence from primary care was a limitation, recognising that early presentation of illness is most likely in that setting, arguably making it the one in which most could be gained from structuring and coding the patient record. Finally, no studies were identified that investigated the influences on choice of codes or text, highlighting the paucity of research into this topic.[13]

2.3 Data storage and format

Data are uploaded to the CPRD electronically from participating surgeries approximately every 4 weeks via the secure NHS intranet. Once uploaded, data are processed into two main storage formats, medcodes or text, depending on the original style of data recording. The alphanumeric Read codes are converted to numeric medcodes in direct 1:1 mapping, such that the two terms are essentially interchangeable.[a] The purpose of this mapping is twofold: first, to reduce the database's size; and secondly, to render the data more amenable to manipulation by information technology packages such as Stata (StataCorp, College Station, Texas, USA) or SPSS (IBM Software, Portsmouth, UK).[14]

Any notes GPs make to supplement the codes are stored in text fields within the CPRD. Hospital letters sent electronically are also imported into text fields; however, letters on paper are scanned to a PDF file, which is attached to the medical record but remains inaccessible to researchers. Storage in PDF format is reducing hand-in-hand with IT developments that encourage paper-based communication to be abandoned in favour of electronic methods. Indeed, the electronic record (incorporating both codes and text) is viewed as the complete legal record.[15]

[a] In this thesis, I retain the term that is appropriate to the setting; therefore, I use 'Read code' when discussing events during consultations with GPs, and 'medcode' in relation to analysis.

2.4 Strengths and limitations

An excellent summary of advances in the utility and use of the GPRD at the time of transition to the CPRD can be found in Williams *et al*[1] A recent profile of the CPRD was published in 2015.[2]

2.4.1 Size

An undeniable strength of the CPRD is its size. As of August 2014, 684 GP practices in England, Northern Ireland, Scotland and Wales were registered with the CPRD, covering approximately 8.8% of the UK population. At that time, the database contained records on 13.56 million patients whose data fulfilled quality criteria specified for research, of whom 5.68 million remained registered with a contributing GP practice.

2.4.2 Probability sample

A further strength relates both to the database's size and the way that health care is provided in the UK. Virtually all UK nationals are registered with a GP, who provides primary care and acts as a gatekeeper to secondary care provided by the National Health Service (mainly hospital-based care). If patients do access secondary care directly, for example by attending the emergency department, this information is shared with the GP. The integration of the systems ensures continuity of the medical record at one point: the GP. It is a fair assumption that virtually all UK citizens have a chance (greater than zero) of

being included in the CPRD, which can therefore be viewed as a ‘probability sample’ of the UK population.

A recent review reported the demographic characteristics of acceptable CPRD patients (as of January 2014), and the subset of those active on 2 July 2013. Of the 4,425,016 active patients, 2,183,161 (49.3%) were men and 2,241,855 (50.7%) were women, reflecting national data. However, the geographical distribution of patients was not even across the regions. For example, 600,824 (13.6%) of patients were in London compared with 29,954 (0.7%) in the East Midlands. Some of this will be accounted for by variations in population density, and some by the requirement of some Clinical Commissioning Groups to use GP systems other than ViSion.[2] Despite these caveats, it is generally accepted that studies using CPRD data may draw inferences from their results to the UK population as a whole, which accounts for much of the database’s appeal.[1,2]

2.4.3 Prospective data collection

As described above (Section 2.2), clinical data are recorded during the consultation, i.e. prospectively. Consequently, exposure and outcome data have limited vulnerability to information bias. In addition, any information transmitted automatically and electronically, such as test results and prescriptions, is free from recording bias.

2.4.4 Data quality

Once uploaded to the CPRD, data are subjected to stringent quality checks for continuity and completeness to ensure that only data of a quality acceptable for research are released. Reports are sent to contributing GP practices highlighting any deficiencies or inconsistencies in the patient records, and subsequent data are refused by the CPRD if the identified shortcomings have not been addressed satisfactorily. A single metric, the ‘up-to-standard date’, is used to indicate the overall quality of data within GP practices. In essence, this is the last date on which the practice was considered to have provided continuous data of a quality high enough for research.[1] One limitation to note is that these quality assurances are only applicable to the coded data, not the text. The internal CPRD metrics are unable to illuminate the uncertainty inherent in the text record, such as spelling and typographical errors or abbreviations whose definition may be either obscure or ambiguous; for example, MVR could be mitral valve repair or mitral valve regurgitation. Therefore, researchers must ensure their search criteria are broad enough to maximise the capture of entries made using shorthand or acronyms. Furthermore, researchers including text data need to devise their own quality checks to quantify the errors associated with any variables they generate from this part of the CPRD.

2.5 Data access and use in research

The CPRD is widely used in epidemiological research, resulting in over 850 research papers to date (www.cprd.com/Bibliography/Researchpapers.asp, accessed 25 September 2015).

Researchers wanting to use CPRD codes as their data source can purchase online access to the most recent version of the primary care datasets via the CPRD GP online data subscription service, aptly known by the acronym CPRD GOLD. Access to powerful query and extraction tools enables these researchers to build their own datasets. An alternative, cheaper, option is to buy *ad hoc* datasets assembled by the CPRD according to the researcher's specified criteria.

Information stored in text fields is not routinely available to researchers, primarily because it may contain information that identifies the patient. This has always been the case; indeed, since the inception of the CPRD, as shown in my literature search, researchers have never routinely requested supplementary text data. Furthermore, a moratorium on collection of CPRD text data was introduced in October 2013, owing to an unspecified governance requirement. The CPRD are working to remove this restriction, so that text data collection can recommence. Until then, the CPRD will provide access to extracts of the text collected before October 2013, in line with a researcher's specified search

terms, although this service is fairly costly owing to the extensive manual checks required to ensure complete anonymisation of the record.[a]

Once the study participants have been identified and their raw data files have been extracted from the CPRD, researchers use statistical software or data management software for further processing and analysis. Codes require a minimum of pre-processing for conversion to variables suitable for analysis.[15] In contrast, manual interpretation and other work-up, all costly and time-consuming, are required before the information stored in text fields can be used in research. The extra cost and analytical complexity of including text data in research present such barriers that the default position is for researchers to restrict their analysis to codes. The ramifications of this data loss are poorly investigated and form the subject of this thesis.

2.6 Other electronic databases

The CPRD is one of three main electronic databases of medical records collected as part of everyday clinical care in the UK. The others are The Health Improvement Network (THIN) and QResearch,[16] which are briefly discussed below.

[a] Note: On 31 March 2016, shortly after my PhD viva, the CPRD stopped providing historical free text (i.e. that collected before 2013).

2.6.1 The Health Improvement Network

The Health Improvement Network (THIN) is a large database of primary care records collected as part of routine clinical care provided by general practices throughout the UK. It was established in 2003 in a collaborative project between In Practice Systems Ltd (INPS, London, UK) and Cegedim Healthcare Software (Boulogne-Billancourt, France), which was later acquired by IMS Health (London, UK). It is primarily used for pharmacoepidemiological research.[17]

THIN shares many of its strengths and limitations (see Section 2.4) with the CPRD, not least because they both collect data using ViSion (INPS). Indeed, a validation study reports that there is considerable overlap between the two databases, with some THIN practices previously contributing, or still currently contributing, data to the CPRD.[17]

Most recent data indicate that 587 general practices currently contribute data to THIN, and that the database holds copies of records for 12.4 million patients, of whom 3.6 million are still active. As this amounts to ~5.7% of the population, the database is generally perceived to be a representative sample of the UK; however, no geographical distribution data for THIN have been published in order to assess the validity of this assumption.[a]

As with the CPRD, data are recorded using a code or in the free text, and the latter are available to researchers at additional cost. THIN is widely used in

[a] See http://www.csdmruk.imshealth.com/our-data/statistics.shtml (accessed 17 May 2016).

research and has generated more than 500 publications to date, with an emphasis on pharmacoepidemiological studies.[a]

The main difference between the CPRD and THIN is one of governance – the CPRD is funded and managed jointly by NHS National Institute for Health Research and the Medicines and Healthcare products Regulatory Agency, whereas THIN is owned by IMS Health and managed via University College London.[b]

2.6.2 QResearch

QResearch is a not-for-profit partnership between University of Nottingham and EMIS Health (Leeds, UK) that was established in 2004.[18] No database profile for QResearch has been published and the following information is taken from either the QResearch website (www.qresearch.org) or from a recent paper published by QResearch co-director, Professor Julia Hippisley-Cox.[19] QResearch currently holds copies of the electronic medical records from approximately 1,000 general practices throughout the UK covering a population of more than 20 million people. Data are collected as part of routine care using

[a] See 'Primary Care Database Publications List' at http://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/publications (accessed 17 May 2016).

[b] See 'THIN Database Research Team' at http://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub (accessed 17 May 2016).

EMIS software.[a] Like the CPRD and THIN, QResearch data include demographic information, diagnoses, prescriptions, referral information as well as laboratory and clinical test results. Diagnoses, symptoms and other clinical information is recorded using Read codes or text.

QResearch shares many of the strengths and limitations of the CPRD (see Section 2.4); however, it is likely that data loss in hidden text is likely to be even greater than in the CPRD because, unlike ViSion, EMIS does not insist on use of Read codes to initiate a medical record. Thus identification of specific diseases and symptoms is more complex than in the CPRD or THIN. It's not clear whether all the record is stored in a free text field, or whether there are separate 'coded' and free text sections. For the purposes of this thesis, there is no available free text section allowing study of missing data.

Various American databases exist, such as Kaiser and UW Medicine. These are created predominately for billing purposes and so have limited applicability for epidemiological research. Similar to QResearch, it is not clear whether there are two tiers of recording.

[a] See 'What is QResearch?' at http://www.qresearch.org/SitePages/What%20Is%20QResearch.aspx (accessed 17 May 2016).

3 Literature review: what is stored as text in the CPRD and what is its value?

A literature review was conducted to identify the type and value of information recorded as text within the CPRD, as this is the first step in clarifying the methodological consequences of omitting text data from analysis. Potential methodological concerns include under-reporting, with attendant underestimation of prevalence. Under-reporting that is not random – for example, being associated with a particular patient subgroup – will lead to bias. These errors are of specific concern for not only the research community but also the policy-makers and health service providers whose decision-making hinges on the quality of the evidence base.

3.1 Literature search criteria

The research question posed was, 'What type of information is stored solely as uncoded free text in the UK-based Clinical Practice Research Datalink (formerly General Practice Research Database), a database of electronic health records?' EBSCOHost was used to identify pertinent literature from the Medline database, including e-Journals. The search terms chosen and number and details of references found are displayed in Table 15.1 (see Appendix 1: Literature search tables): 51 papers of potential relevance to the research question were identified. After review, 10 published papers were retained for the final review. Reference lists within the retained publications were checked, which identified one relevant conference proceedings report. Citations of all

selected publications were found using Web of Science, but identified no further papers for inclusion in this review. Reasons for exclusion are given in Table 15.2.

3.2 Findings from use of text in validation studies

In systematic reviews, Herrett *et al*[20] and Khan *et al* [21] reported on the value of text data as a tool for 'validating' clinical data recorded elsewhere in the CPRD. Nicholson *et al*[15] defined validation in this context to mean not only how accurately the selected code reflected the GP's thinking at the time but also whether the correct diagnosis was made. Internal validation, against the patient's medical record or by separate GP questionnaire, is a standard way of checking agreement between a code and the GP's opinion. Confirming whether the correct diagnosis was made involves external validation against a gold standard, often the opinion of a hospital specialist. Increasing integration of the entire medical record (including communications from secondary care) into text fields illustrates the significance of this part of the CPRD to validation, and hence to research.

Seven validation studies were identified whose results suggest the type of information that is stored solely as text in the CPRD. Their findings also illuminate how to optimise access to, and use of, information in text fields and hint at the ramifications of ignoring this part of the CPRD in terms of identifying cases, outcomes and exposures.

3.2.1 Pregnancy and pregnancy outcome validation studies

The CPRD is an excellent database for examining pregnancies and pregnancy outcomes, because its research information requirements were designed to enable the data to be captured in sufficient detail for use in drug safety studies.[3] The current CPRD's data recording guidelines for ViSion, published in 2004, recommend that Read codes are used to record live births, spontaneous abortions and terminations in the mother's record.[3] Foetal abnormalities and serious congenital malformations should also be noted in the mother's record. This is to protect against data loss in the event of a prolonged hospital stay or early death, as a result of which the infant may never register with the mother's GP practice. The guidelines stress the importance of making it clear that such outcomes relate to the foetus/infant and not the mother. To achieve this, they suggest input of a 'general' Read code indicating contact with the health service for something other than illness (often denoted with the prefix [V]), and using the text field to note the details of the anomaly. Therefore, text has the potential to be a rich source of information for studies of adverse pregnancy outcomes. Three studies have examined the CPRD's utility as a validation tool, as described below.

First, Wurst *et al*[22] carried out a validation study of CPRD data recorded between January 1992 and February 2005 for three specific congenital heart defects: ventral septal defect, coarctation of the aorta and tetralogy of Fallot. Case records are commonly validated by questionnaire in conventional research projects, in an expensive and time-consuming procedure. The authors

sought to find whether information in the mothers' and infants' text fields in the CPRD could be extracted and used for validation in place of questionnaires. A thesaurus of codes pertaining to the congenital heart defects was drawn up, reflecting codes used by GPs during the time of the study. Read codes and their forerunner Oxford Medical Information System (OXMIS) codes, which were used in VAMP, identified 24 cases of coarctation of the aorta, 72 cases of tetralogy of Fallot and 373 cases of ventral septal defect in the CPRD. Of these, 104 were randomly selected for inclusion in the study. Text fields in the infant record paired with the first occurrence of a Read or OXMIS code were searched for keywords relating to congenital heart defects, such as 'congen*', 'heart', 'Fallot', 'tetra', 'cardiac', 'defect', 'anomal*' and 'septal'. All text fields in the maternal record were searched for the same terms in the year before and 2 years following the infant's delivery date. Outcome measures included concordance between information in text fields and that obtained from a practitioner questionnaire. Text records were available for only about 50% of the infants in the study. Nevertheless, concordance with questionnaires was high (ranging from 92% to 100% on the various measures), leading the authors to conclude that infant text fields have potential for case report validation. Text from the maternal record yielded no information that was not otherwise available through the medical record or the infants' text fields. Indeed, only 31% of mothers' records contained any information at all about the infant's heart defect. This is perhaps unsurprising, given that the guidelines recommending use of mothers' text fields to record details of anomalies were published in 2004, near the end of the study period. The authors recommended that future

studies should examine infant text fields paired with every occurrence of the Read/OXMIS code rather than just the first, to maximise the amount of information retrieved and minimise the need for questionnaire validation. However, in light of the updated recommendations in force since 2004, it would be unwise not to include the mothers' text records in future studies as well.

Secondly, Devine *et al*[23] developed an algorithm to identify pregnancies within the CPRD and subsequent outcomes over the period 1 January 1987 to 31 December 2006. A thesaurus of pregnancy codes (Read and OXMIS codes) was compiled and each code was categorised in relation to an end-of-pregnancy event or pregnancy care. Two tools were used to validate the algorithm. In the first, live birth and miscarriage counts generated from the algorithm were compared with data held in the additional clinical details maternity (ACDM) file, which contains information on completed births and miscarriages in the CPRD. In the second, text fields recorded 2 weeks either side of the end-of-pregnancy date were searched for a list of terms chosen to optimise the identification of a wide range of possible pregnancy outcomes. The presence of one of these terms was taken as validation of the algorithm. Concordance between the algorithm and information in text fields was reported in terms of percentage agreement and kappa score. Both validation tools were consistent with the algorithm for the outcome of live births. However, for spontaneous abortion and miscarriage, while the ACDM file did validate the algorithm, the text did not (percentage agreement 77%, kappa score 0.36). Similarly low kappa scores were found for other pregnancy outcomes, such as

elective termination and multiple births. The authors suggested that use of less specific search terms for live births compared with other possible pregnancy outcomes explained their results. They concluded that future studies should examine the entire text field rather than search for particular terms to remove dependence on search term specificity. It is worth considering whether the choice of study period can shed further light on the authors' results. The chosen study period (1987-2006) coincides with the initiation and early development of VAMP, during which time recommendations regarding use of codes and text were being refined. It would be interesting to re-examine the performance of the validation tools during a period of consistent recording practice. In mitigation, the number of pregnancies identified early on was relatively small (271 in 1987, 15,750 in 1990) compared with later on (43,408 in 2006).

Thirdly, Charlton *et al*[24] studied the utility of the CPRD in the surveillance of pregnancy outcomes to identify potential teratogens. The study population was derived from women in the CPRD with a live pregnancy outcome between 1 January 1990 and 31 December 2006 and a medcode for a diagnosis of epilepsy, seizure or convulsion. The offspring of these mothers were identified and mother-baby pairs included in the study sample (*n* = 3,869) if the baby remained registered within the CPRD for 3 months after birth. Using a thesaurus of pertinent medcodes, 188 potentially major congenital malformations in 161 children were identified. Verification of the congenital malformation was carried out by manual review of a photocopy of the child's complete medical record (*n* = 109 malformations) or, where this was not

available, the entire text (*n* = 77). The medical record verified the coding of 100 of the 109 (91.7%) congenital malformations, whereas text performed less well, verifying 60 out of 77 (77.9%). The authors reported that text was less reliable than the medical record as a verification tool and presumed this was partly because it omits paper communications, which are not available to researchers. As noted above, this is likely to reduce as a problem in the long term, because increased use of electronic communication between other healthcare providers and GPs will enhance the level of detail recorded in searchable text fields. However, the study was limited by not examining the mothers' text records, as this may have been a valuable source of information about congenital malformations.

3.2.2 Other validation studies

In response to concerns over the completeness and accuracy of information held in cancer registries, Boggon *et al*[25] carried out a validation study to compare information held in the National Cancer Data Repository (NCDR) and the CPRD. Concordance between the two datasets was taken as confirmation of cancer. If cases were identified on the NCDR but not the CPRD, text was searched for a thesaurus of terms validated separately by manual inspection of a small sample of CPRD records. Overall the level of concordance between the two databases was high. Across all cancers combined, the CPRD identified 2.1% more cases than the NCDR; however, for colorectal, lung, urinary tract and pancreatic cancers the converse was true. Of the 5,676 cases identified in the NCDR, 5,335 (94%) were recorded in the CPRD, 624 of which were stored

in text fields only, amounting to 11.0% of the NCDR figure. The authors considered a number of possible reasons for the apparent disagreement between the data, including the fact that the CPRD and NCDR use different coding thesauri, and the stage of disease at diagnosis. For lung and pancreatic cancers in particular, which have a poor prognosis, some patients present as emergency admissions to hospital and are diagnosed with cancer shortly before their death. It is feasible that such diagnoses are never recorded in the patient's primary care record.

Boggon *et al*'s study indicated not only the importance of verifying diagnostic code thesauri, but also the potential use of text for case ascertainment, particularly for the two cancers investigated as part of my PhD, bladder and pancreatic cancer.

Close *et al*[26] estimated the relative risk of renal failure or impairment in patients with bipolar disorder according to age, gender and exposure to lithium. The study also attempted to validate recording of lithium use in Prescription Pricing Authority (PPA) codes within the CPRD. Internal validation procedures included examination of the complete text field, although this was only available for 44 of the 77 (57%) confirmed cases of renal failure. Of these 44 patients, all 28 who had PPA codes for lithium were validated by text. Lithium exposure duration was confirmed for 22 of these 28 patients; for the remaining 6 patients, duration was reportedly underestimated by between 4 and 32 years (mean 12.3 years, SD 10.3 years). However, the large size of the standard deviation relative to the mean suggests that the data were not normally distributed and that a non-

parametric measure of central tendency would have been more appropriate. Letters from secondary care for 16 patients who did not have PPA codes for lithium indicated that they had indeed been prescribed the drug. The authors concluded that exposure to lithium and potentially other medications may be underestimated in studies using CPRD data, especially if uncoded data are ignored. The study can be criticised not only for its choice of descriptive statistics, but also for its approach to determining both renal failure and lithium use. For renal failure, diagnostic and referral codes were used, but laboratory test results of kidney function, objective markers of renal failure, were omitted. For lithium use, using the therapy file rather than PPA codes would have captured all prescriptions originating in primary care and, in theory although maybe not in practice, those in secondary care. This is important because lithium is a specialist drug, and it is plausible that not all prescriptions were issued by the GP. Therefore, one could argue that the authors did not choose the optimal validation tool available to them.

Thomas *et al*[27] reported on the validation of suicide and self-harm recording in the CPRD between 1 January 1998 and 31 December 2010 by comparison with Office for National Statistics (ONS) and Health Episode Statistics (HES), respectively. Coded CPRD data failed to identify 1,670 (74%) out of 2,260 ONS-confirmed suicides between these dates. Using wild cards to allow for variation in terms, they searched the text fields for pertinent phrases, for example ‘suicid*’, and picked up an additional 179 completed suicides, accounting for 10.7% of the missed cases. Coded CPRD data on self-harm failed to identify

622 patients admitted to hospital (HES data) with self-harm in 2010. Searching the text for phrases such as 'overdose', 'self*harm*' and 'self*injur*' identified 101 (16.2%) of these missed cases of non-fatal self-harm. Thomas *et al* concluded that CPRD codes underestimate the incidence of both completed suicide and self-harm and that information stored in the text accounts for only part of this under-recording. Initially, one might be unsurprised to discover that the CPRD under-reports suicide, an event that generally occurs outside a GP's surgery. However, this finding should be considered in terms of the current recording guidelines for ViSion[3] (while recognising that they only came into force in the middle of study period). The CPRD emphasise the need to record the date, fact and cause of death, as all are extremely important outcomes for research; indeed, it is a contractual requirement for practices to notify their Health Authority of a patient's death. Consequently, the CPRD designed a Death Administrative Management Plan specifically to collect this information using Read codes. The guidelines list specific codes that might be selected and do not specify any circumstances where text should be used instead. It would be valuable to know whether the fact of death was also under-recorded in the CPRD for those patients whose suicide was not identified, but regrettably these data were not reported.

Finally, Shah *et al*[28] analysed text fields to validate an algorithm for identifying cause of death in a random sample of 3,310 patients who died in 2001. The algorithm performed with a positive predictive value of 98.4% (95% CI: 97.2–99.2) and sensitivity of 92.9% (95% CI: 90.8–94.7). Cause of death was

recorded solely as text in 19.4% of a random sample of 3,310 patients registered in the CPRD who died in 2001, suggesting that coded data alone under-report this outcome. Again, it would be valuable to know whether the fact of death was similarly under-reported.

3.3 Findings from studies into automated text processing

As indicated by the validation studies discussed above, valuable information can be obtained from studying the text fields and progress is steadily being made to overcome the aforementioned barriers to routine use of these data. Indeed, techniques are available to automate the interpretation of text, ranging from simple searches for key terms, through computer-based algorithms, to natural language processing (NLP). NLP is roughly defined as intelligent processing of human language by computers; it has application in both information extraction and the automated interpretation of text.[29] There is understandable interest in applying NLP techniques to research using electronic health records. The Patient Records Enhancement Programme (PREP), funded by the Wellcome Trust and based at the Universities of Brighton and Sussex, aims to enhance access to text fields and is working closely with the CPRD. The findings of studies published as part of this programme are discussed below, in terms of information 'hidden' in the text.

3.3.1 The PREP studies

Tate *et al*[30] focused on how accurately coded diagnosis dates matched the actual dates on which the diagnoses were received. Participants were 344

women with a coded diagnosis of ovarian cancer recorded between 1 June 2002 and 31 May 2007. The group obtained the entire text record for these patients recorded in the year prior to their diagnosis. The results suggested that omitting to examine the text may mislead researchers into thinking that a diagnosis was made later than it actually was: for 22% of 344 women in the study, evidence of the diagnosis was recorded in the text before the diagnostic code was assigned. The median time difference was 24 days (interquartile range: 8–67 days), but for 34 (10%) of patients, the diagnosis was recorded in the text more than 4 weeks before it was coded. The authors acknowledged the limitation that they did not investigate whether ignoring the text causes information to be missed completely (i.e. false-negative misclassification), because they did not identify patients whose diagnosis was recorded solely as text.

Nicholson *et al*[31] picked up on the aforementioned limitation and analysed text fields in an attempt to maximise the probability of identifying all patients with a diagnosis of rheumatoid arthritis within the CPRD. The authors proposed that evidence of the disease should be sought from three sources: level 1, diagnostic codes for rheumatoid arthritis; level 2, codes from other domains, such as symptoms or test results; and level 3, text. The amount of information recorded solely in the text was not reported; nevertheless, the authors concluded that using codes alone for case definition carried the potential both to miss patients diagnosed with the disease and to assign an incorrect date on which the diagnosis was made.

Ford *et al*[32] examined CPRD medcodes and text on 6,387 patients aged 30 years and older who had a first coded diagnosis of rheumatoid arthritis between 2005 and 2008. They searched all text recorded in the year before the medcode was entered and found key words indicating a diagnosis of rheumatoid arthritis in 29% of these patients. Additionally, text contained key words suggesting a diagnosis of inflammatory arthritis in 14% of patients but only 706 (11%) actually had a diagnostic code for the disease. Codes for synovitis were recorded for just 179 (3%) of patients, but a search of the text revealed that an additional 1,168 (18%) of the patients had keywords related to this condition. Finally, a positive test result for rheumatoid factor was found in text alone in 13% of the patients. The text note was most likely to be associated with medcodes for letters and communications. The authors acknowledged that a major limitation of their work is that they did not identify negation of their key words; therefore, a proportion of the uncoded records may denote the absence of a symptom of inflammatory arthritis, or its presentation by someone other than the patient. The authors conceded that they may have overestimated the extent of information hidden to studies restricted to codes.

As noted in a conference proceedings report, Koeling *et al*[33] analysed the same dataset as Tate *et al*[30] and searched the text fields (GP supplementary notes and hospital letters) of all 344 patients in the year prior to their diagnosis of ovarian cancer for references to five commonly occurring symptoms. Negation of symptom reporting was identified using an algorithm. The incidence of each symptom was increased markedly by inclusion of information recorded in the

text: abdominal pain rose from 147/344 (43%) to 208/344 (60%); urogenitary problems, 87 (25%) to 140 (41%); abdominal distension/bloating, 86 (25%) to 190 (55%); constipation, 57 (17%) to 127 (37%); and diarrhoea, 28 (8%) to 87 (25%). To the best of my knowledge this is the only CPRD study to have examined symptom recording restricted to text fields.

3.4 Conclusions

This review confirms that few studies include information recorded in text in the CPRD, with the majority restricting their analysis to coded data. Therefore, even if the effects of attendant data loss are small, the impact will be pervasive. Indirect evidence from validation studies suggested that ignoring text fields incurs data loss, leading most notably to inaccuracies in patient outcomes and case ascertainment. Validation studies revealed nothing about loss of symptom information and the evidence regarding loss of data on drug exposure was rather weak. Nevertheless, it is apparent that valuable information can be obtained from studying the text fields and that CPRD codes alone do not reflect the complete medical record. The PREP studies provided stronger evidence of data loss, and suggested that restricting analysis to coded data will lead to errors in diagnosis date and case ascertainment as well as underestimation of symptom prevalence.

To the best of my knowledge, no studies have attempted to quantify any bias introduced by this data loss. This is important not least because CPRD data are increasingly used in primary care research studies, whose findings have the

potential to influence public health policy and practice within the National Health Service. As this review highlights, an unknown quantity of potentially valuable information is held solely in text format within the CPRD, with barriers to access meaning that it is generally ignored by research. Rapid advances in the automated interpretation of text are being made, such that the inclusion of text into research studies may become routine in the future, but only if governance restrictions on the collection of such data are lifted (see Section 2.5). However, until such time arrives, it is important to investigate and quantify the impact of ignoring this section of the CPRD on research outcomes. These topics form the basis of my PhD studies.

4 Using symptoms of cancer to study the impact of data loss in 'hidden' text

4.1 Introduction

Cancer is a significant source of morbidity and mortality in the UK. For example, the latest data summarised by Cancer Research UK indicate that, in 2012, over 338,000 people in the UK were diagnosed with a form of the disease, and that there were approximately 162,000 cancer deaths.[34] Survival of most adult cancers in the UK has generally been poorer than that in comparable western European countries.[35] This observation prompted a number of UK Government policies, including *The NHS Cancer Plan*[36] in 2010 and its successor *The National Cancer Strategy*[37] in 2011. The clinical evidence to support policy decisions has been provided by a number of research programmes, including the National Awareness and Early Diagnosis Initiative (NAEDI) and the Discovery Programme.

Discovery was a UK-based programme funded by the National Institute for Health Research whose aim was to expedite the diagnosis of cancer in symptomatic patients who present in primary care. The Universities of Bristol, Cambridge, Bangor, Durham, Oxford and Exeter as well as Bristol NHS Clinical Commissioning Group all contributed to the programme, which officially ended in June 2015. Theme 2 of the Discovery Programme analysed CPRD data to model the presentation of cancer in primary care; therefore, it was selected as

an ideal forum for investigating the potential impact of symptom data loss in 'hidden' text.

4.2 Theme 2, the CAPER Studies

In Theme 2, the Cancer Prediction in Exeter (CAPER) Studies determined how cancer – 13 sites to be studied in total – typically presents in primary care. Observational studies for each cancer were conducted with data from the CPRD (see review in Section 2). Using a variant of the case–control study design, the main clinical features of cancer presentation were identified and their associated risk of cancer quantified, singly and in pairs. The main outputs of this work are risk assessment tools, which describe the risk of cancer with symptom combinations.

An evaluation of risk assessment tools was conducted in a cohort study of 614 GPs from 165 practices in England. The results suggested that use of risk assessment tools was accompanied by increased diagnostic activity and additional cancer diagnoses. However, the study was not designed to determine whether these changes were wholly attributable to implementation of the risk assessment tool.[38]

A review of risk prediction tools for cancer in primary care was published recently. It highlighted the uncertainty that remains over their clinical utility, and called for further research into their implementation in primary care, so that their use maximises benefits while minimising harm.[39] This requires knowledge of the risk assessment tool's accuracy to predict the chances of cancer in symptomatic

patients. My study is designed specifically to address one aspect of this – the introduction of errors associated with restricting analysis to clinical information recorded by GPs using Read codes, and ignoring anything noted in the text.

4.3 Selection of cancer and symptoms to be studied

At the time of writing, the CAPER Studies had published their findings on the main features of nine cancers in primary care – bladder,[4] pancreatic,[5] kidney,[40] oesophago-gastric,[41] breast[42] and uterine[43] cancers, as well as myeloma[44] and non-Hodgkin[45] and Hodgkin[46] lymphomas.

Bladder and pancreatic cancer were chosen to be the exemplars, because they share a common symptom, abdominal pain, thus permitting comparison of its recording between cancers. In addition, each of these cancers has a characteristic headline or 'alarm' feature. Text recording was examined for three features; namely, haematuria (the alarm symptom for bladder cancer), jaundice (the alarm feature for pancreatic cancer) and abdominal pain. Additional reasons for their selection included the fact that they are not part of the Quality and Outcomes Framework initiative; therefore, their recording is unlikely to be influenced by financial incentives.[47] Finally, uncertainty around their presence and severity varies, which may affect GPs' recording style. Abdominal pain has many causes and is assessed by both clinical examination and subjective patient reporting. In contrast, jaundice and visible haematuria are determined objectively and both arouse a strong clinical suspicion of serious pathology.

The cancers and symptoms chosen for this study are now discussed in detail, to give sufficient context for interpretation of the clinical and methodological findings of the study.

4.4 The cancers

4.4.1 Pancreatic cancer

4.4.1.1 The pancreas

The pancreas is a soft, lobulated gland that lies between the posterior abdominal wall and the peritoneum at the level of the lumbar spine (L1–L2). The organ is 12–15 cm long and, shaped like the letter J, consists of a head, neck, body and tail. Its head sits in a curved loop of duodenum, from which point the body extends superiorly and to the left. Its tail – forming the bulk of the pancreas – lies adjacent to the spleen, close to the posterior abdominal wall. The pancreas functions as an exocrine gland, secreting digestive enzymes, and as an endocrine gland, secreting the hormones insulin and glucagon.[48]

4.4.1.1.1 Exocrine function

The exocrine portion makes up 98% of the pancreas. It produces enzymes essential for the breakdown of carbohydrates, proteins and fats ingested as food into smaller molecules that can be absorbed from the small intestine into the blood stream.

Enzyme synthesis occurs in acinar cells in the pancreatic tail. In brief, three main types of enzyme are produced: protease (to digest protein), lipase (to digest fat) and amylase (to digest carbohydrate).

Acinar cells are clustered into lobules around a central lumen that forms the beginning of a ductule. The ductules draining these lobules form a network of ducts that feed into the main pancreatic duct (duct of Wirsung) that runs from the tail of the pancreas to its head. Cuboidal epithelial cells lining the pancreatic ducts release a bicarbonate-rich fluid, which forms an alkaline pancreatic juice in which the digestive enzymes flow along the ductule system. The main pancreatic duct joins the common bile duct, which originates in the gallbladder, and together they drain into the duodenum at the ampulla of Vater. Thus, there is a clear delivery pathway from the place of digestive enzyme production, the acinar cell, to the site of action in the duodenum.[48,49]

4.4.1.1.2 Endocrine function

The endocrine portion makes up 2% of the pancreas, in the form of cells called Islets of Langerhans. These cells secrete hormones, most of which are essential for the control of blood glucose; for example, insulin and glucagon. Islets cells are highly vascularised; even though they make up just 2% of all pancreatic tissue, they receive a disproportionate 10–15% of the organ's overall blood supply. This facilitates the release of their hormones directly into the blood supply – the distribution of endocrine pancreatic hormones does not take place via the exocrine pancreatic duct system.

4.4.1.2 Definition of pancreatic cancer

Malignant neoplasms can arise in the head, neck, body or tail of the pancreas. According to the International Classification of Disease, pancreatic cancer falls under the category C25.

4.4.1.3 Risk factors

Parkin *et al*[50] estimated that 28.7% and 12.2% of the incident pancreatic cancer cases in the UK in 2010 were due to tobacco smoking and obesity, respectively. In a review article, Hidalgo cites evidence that the risk of pancreatic cancer in smokers is 2.5 to 3.6 times that in non-smokers, with the risk increasing the greater the exposure.[51]

Aside from these lifestyle factors, there is some evidence to suggest a familial cause in some cases. For example, in a collaborative case–control study, the Pancreatic Cancer Cohort Consortium found evidence of a moderate association between a family history of pancreatic cancer and risk of developing the malignancy (odds ratio 1.76, 95% CI: 1.19–2.61). The association may arise from shared genetic risk factors, such as inherited mutations, although these are rare. The association may also arise from shared environmental risk factors.[52]

Other lifestyle factors, such as alcohol, red meat consumption and exposure to radiation, have been investigated but the evidence for their association with pancreatic cancer is limited.[53]

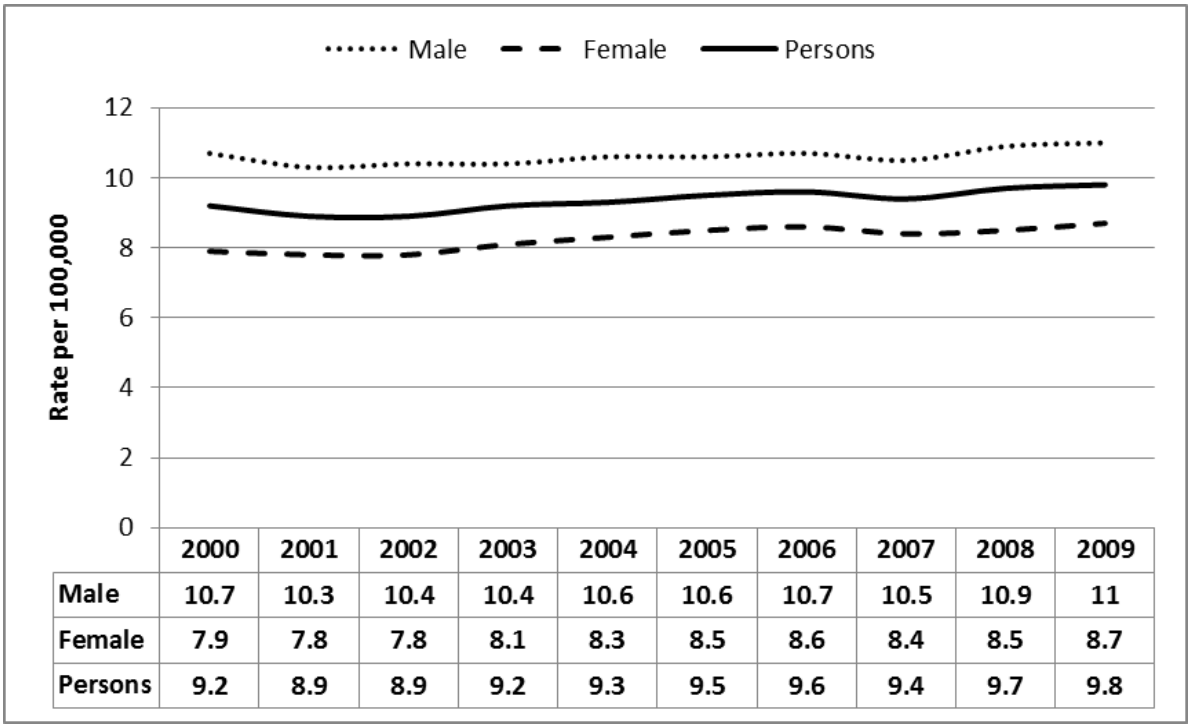
4.4.1.4 Incidence

The latest statistics summarised by Cancer Research UK report that 8,773 people (4,328 new cases in men; 4,445 in women) were diagnosed with pancreatic cancer in the UK in 2011. These figures equate to age-standardised incidence rates of 10.8 (95% CI: 10.4–11.1) per 100,000 men and 8.7 (95% CI: 8.5–9.0) per 100,000 women. In terms of ranking, pancreatic cancer is the 13th most common cancer in men and the 9th most common cancer in women.[53]

The incidence of pancreatic cancer rises with age. Current incidence rates are low and similar in men and women below the age of 40 years. The incidence in men and in women rises sharply as age increases above 50 years; for example, from 5 to 9.2 per 100,000 men and from 3.3 to 7.3 per 100,000 women in the age bands 45–49 years and 50–54 years, respectively. Incidence then increases steadily, to peak in those aged 85 years or older, at 109.4 per 100,000 men and 92.6 per 100,000 women.[53]

Aggregate incidence data for pancreatic cancer in each year of the recruitment period of my study are plotted in Figure 4.1. The incidence in men was fairly stable throughout the study period at between 10 and 11 cases per 100,000 men. In women the incidence fluctuated above and below 8 per 100,000 women between 2000 and 2004, after which it tended to rise to reach a value of 8.7 per 100,000 women in 2009.[53]

Figure 4.1 European age-standardised incidence rates for pancreatic cancer per 100,000 population; UK data 2000–2009



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Male	10.7	10.3	10.4	10.4	10.6	10.6	10.7	10.5	10.9	11
Female	7.9	7.8	7.8	8.1	8.3	8.5	8.6	8.4	8.5	8.7
Persons	9.2	8.9	8.9	9.2	9.3	9.5	9.6	9.4	9.7	9.8

Note: Adapted from a graph prepared by Cancer Research UK

Original data sources:

1. Office for National Statistics. Cancer Statistics: Registrations Series MB1. http://www.statistics.gov.uk/statbase/Product.asp?vlnk= 8843.

2. Welsh Cancer Intelligence and Surveillance Unit. http://www.wcisu.wales.nhs.uk.

3. Information Services Division Scotland. Cancer Information Programme. www.isdscotland.org/cancer.

4.4.1.5 Diagnosis

Pancreatic cancer only becomes clinically apparent after the malignancy has advanced enough to invade surrounding organs, or has metastasised. Unfortunately, there is no effective screening tool for pancreatic cancer, meaning that it is usually diagnosed when patients become symptomatic, in other words at an advanced stage of the disease.[54]

In a recent meta-analysis of the association between symptoms and pancreatic cancer, Schmidt-Hansen *et al*[55] concluded that the only high-risk feature of pancreatic cancer in primary care was jaundice. Meta-analysis of the positive predictive value of jaundice for pancreatic cancer was not possible owing to the lack of suitably designed studies.

Stapley *et al* – whose CAPER study I extend in my PhD – reported a positive predictive value of jaundice for pancreatic cancer of 12.9% (95% CI: 7.9–7.1%) in patients aged over 40 years.[5] Based on Stapley *et al*'s study, The National Institute for Health and Care Excellence (NICE) recommend referral of patients who are 40 years or older and who have jaundice for investigation for suspected pancreatic cancer. The guidelines also recommend GPs to consider referral in patients who are 60 years or older with weight loss plus any of the following symptoms: diarrhoea, back pain, abdominal pain, nausea, vomiting, new-onset diabetes[a] or constipation.[9]

The National Cancer Intelligence Network[56] identified routes to diagnosis for the 57,566 pancreatic cancer diagnoses made in the period 2006 to 2010, at a time when there were no specific recommendations for referral if GPs suspected a patient may have pancreatic cancer.[57] The data indicate that 27,056 (47%, 95% CI: 47–48%) were diagnosed following an emergency presentation. A further

[a] New-onset diabetes was determined as the first ever occurrence of a diagnostic Read code for diabetes in the patient's medical record, or as the first ever occurrence of a raised blood glucose level above the local laboratory's normal range.

12,089 (21%, 95% CI: 20–21%) were diagnosed following referral by the GP, but not under the auspices of the two-week-wait rule. This latter pathway, which was introduced in December 2000, led to the diagnosis of 8,060 (14%, 95% CI: 14–14%) patients with pancreatic cancer.

These results are supported by Lyratzopoulos *et al*'s[58] findings from the 2010 National Cancer Experience Survey in England. Of the 24 cancers studied, pancreatic cancer was ranked the second highest in terms of percentage of patients who consulted their GP three or more times before a referral was made (193 of 467, 41.3%).[a]

Pancreatic cancer is commonly diagnosed using computed tomography. The majority of tumours originate in the exocrine pancreas, and are known as ductal adenocarcinoma.[54] Tumours originating in the endocrine pancreas – neuroendocrine tumours such as insulinomas and glucagonomas – are relatively rare, and most are benign.[53]

4.4.1.6 Stage at diagnosis

Once diagnosed, cancer is staged to assess the size of the original tumour and whether it has spread in the body – this is generally the case for all cancer types, not just pancreatic cancer. A commonly used staging system is TNM (tumour, node, metastases):

[a] Only multiple myeloma was greater, with 939 of 1,854 (50.6%) of patients consulting their GP three or more times before referral.

- **T** indicates the size of the cancer and how far it has spread into nearby tissue. T can take a value of between 1 (small) and 4 (large)
- **N** indicates whether the cancer has spread to the lymph nodes. N can take a value between 0 (no lymph node involvement) and 3 (lots of lymph nodes are involved)
- **M** indicates whether the cancer has spread to another part of the body. M can be 0 (no spread) or 1 (the cancer has spread)

Once this TNM stage has been established, cancer diagnoses are assigned a general stage of between I and IV (Table 4.1).

Table 4.1 Staging of cancers

Stage	**Description**
0	Carcinoma *in situ*
I	Relatively small cancers, contained within the organ of origin
II	Cancers larger than stage I cancer, but which have not started to spread to surrounding tissue. Sometimes cancer cells may have spread into lymph nodes close to the tumour (depending on the cancer type)
III	Relatively large cancers that may have started to spread to surrounding tissues. Cancer cells have usually spread to lymph nodes in the area
IV	The cancer has spread from its organ of origin to another organ within the body – also known as secondary or metastatic cancer

The National Cancer Intelligence Network extracted staging data from the English National Cancer Registration Service for cancers diagnosed in 2013[a]; however, data specific to pancreatic cancer were not reported.

In a US-based study to validate the staging of pancreatic cancer, Bilimoria *et al*[59] identified the stage at diagnosis of 121,713 patients in the US National Cancer Data Base registered in the period 1992–1998. Of these patients, 67,192 (55.2%) were diagnosed at stage IV, and 15,831 (13.0%) at stage III.

As described above (see Section 4.4.1.5), in the period of 2006–2010 in the UK, nearly half of all pancreatic cancers were diagnosed following emergency admission. This reflects the fact that pancreatic cancer remains asymptomatic for a long while, and only becomes symptomatic once it has spread to other regions of the body. Even then, the initial symptoms are vague and common to a number of diseases, such that by the time the symptoms are thought to be attributable to pancreatic cancer, the disease has progressed to a late stage.[54]

4.4.1.7 Mortality

The Office for National Statistics report that 7,546 deaths attributable to pancreatic cancer (C25) were registered in England and Wales in 2013, accounting for 5% of the 145,344 deaths from cancer registered that year. Gender-specific data indicate that similar numbers of men (*n* = 3,767) and women (*n* = 3,779) died from pancreatic cancer in 2013, equating to just under

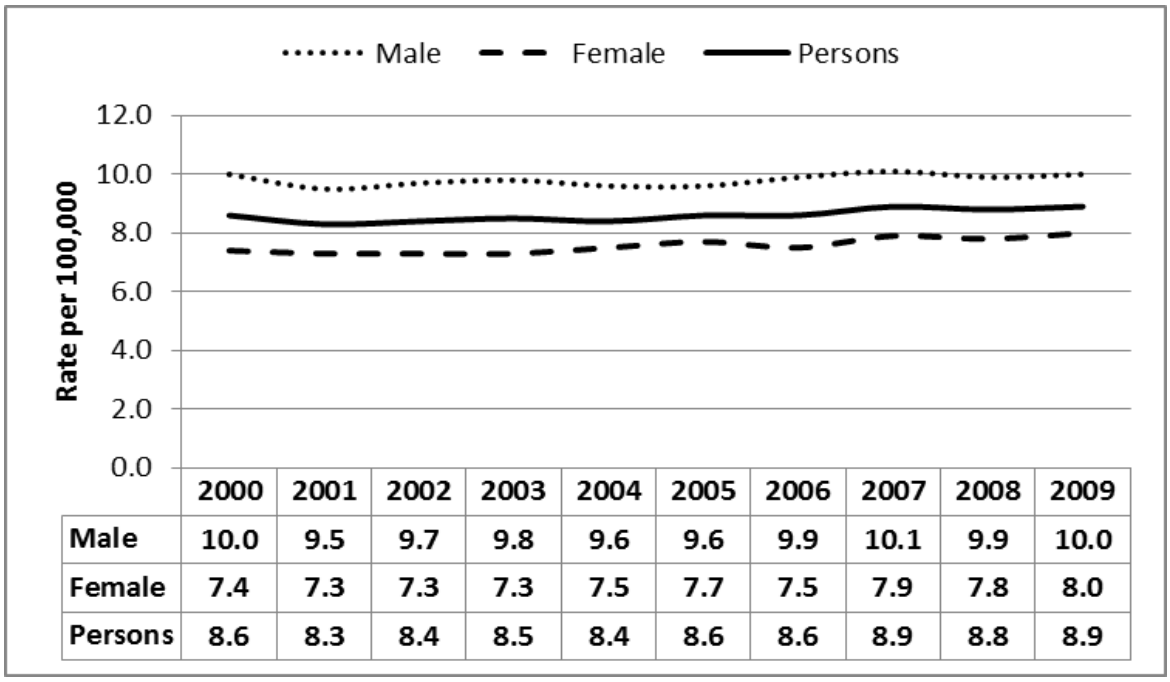
[a] See http://www.ncin.org.uk/publications/survival_by_stage, accessed 12 November 2015.

5% of the 76,962 cancer deaths in men, and to just over 5% of the 68,382 cancer deaths in women. The majority (6,629/7,546= 88%) of all pancreatic cancer deaths occurred in patients who were 60 years of age or older.[60]

Pancreatic cancer was the fifth most common cancer cause of death in men, and the fourth most common cancer cause of death in women in the UK in the previous year, 2012. Correspondingly, the European age-standardised mortality rate from pancreatic cancer in the UK in 2012 was 10.2 (95% CI: 9.9–10.6) per 100,000 men and 8.2 (95% CI: 7.9–8.4) per 100,000 women.[53]

Aggregate mortality data for pancreatic cancer in each year of the recruitment period of my study are plotted in Figure 4.2. The pattern was similar to that of incidence, because survival (see Section 4.4.1.8) is so poor. Mortality in men was fairly steady, at between 9.5 and 10.0 per 100,000 men, whereas in women it was stable between 2000 and 2003, at about 7.3 per 100,000 women, after which it rose to reach 8.0 per 100,000 women in 2009.[53]

Figure 4.2 European age-standardised mortality rates for pancreatic cancer (C25) per 100,000 population; UK data 2000–2009



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Male	10.0	9.5	9.7	9.8	9.6	9.6	9.9	10.1	9.9	10.0
Female	7.4	7.3	7.3	7.3	7.5	7.7	7.5	7.9	7.8	8.0
Persons	8.6	8.3	8.4	8.5	8.4	8.6	8.6	8.9	8.8	8.9

Note: adapted from a graph prepared by Cancer Research UK

Original data sources:

1. Office for National Statistics, Mortality Statistics: Deaths registered in England and Wales: http://www.ons.gov.uk/ons/search/index.html?newquery= series+dr

2. General Register Office for Scotland, Deaths Time Series Data, Deaths in Scotland: http://www.gro-scotland.gov.uk/statistics/theme/vital-events/deaths/time-series.html

3. Northern Ireland Statistics and Research Agency, Deaths by cause: http://www.nisra.gov.uk/demography/default.asp14.htm

4.4.1.8 Survival

The terminology of survival data is complex and requires clear definition. I follow the definitions used by Cancer Research UK – the source of the data discussed in this section. Namely, '*net survival* estimates the number of people who survive *their* cancer rather than calculating the number of people diagnosed with cancer who are still alive. In other words, it is the survival of cancer patients

after taking into account the background mortality that they would have experienced if they had not had cancer.' In contrast, '*relative survival* compares the survival of individuals with cancer to those in the general population. Ideally it would be to those without cancer, but this baseline is difficult to obtain. It is similar to the probability of survival from cancer without including any other cause of death.'[61]

The 1-year, age-standardised, net survival rates for men and women diagnosed with pancreatic cancer in England and Wales during 2010–2011 are reported in Table 4.2. In this time period, survival at 1 year was similar in men (21.6%, 95% CI: 20.6–22.5%) and women (20.1%, 95% CI: 19.1–21.2%). Survival of patients diagnosed in 2010–2011 was predicted to be very poor in men and women at both 5 and 10 years (Table 4.2); indeed, pancreatic cancer ranks the lowest of all cancers in terms of 10-year survival.[53]

Table 4.2 Pancreatic cancer, age-standardised 1-, 5- and 10-year net survival, in adults aged 15–99 years), England and Wales, 2010-2011

	Net survival (%, 95% CI) at		
	1 Year	**5 Years [a]**	**10 Years [a]**
Men	21.6 (20.6–22.5)	3.5 (1.7–6.5)	1.1 (0.1–6.5)
Women	20.1 (19.1–21.2)	3.1 (1.3–6.2)	1.1 (0.1–6.4)

[a] Predicted using an excess hazards model.

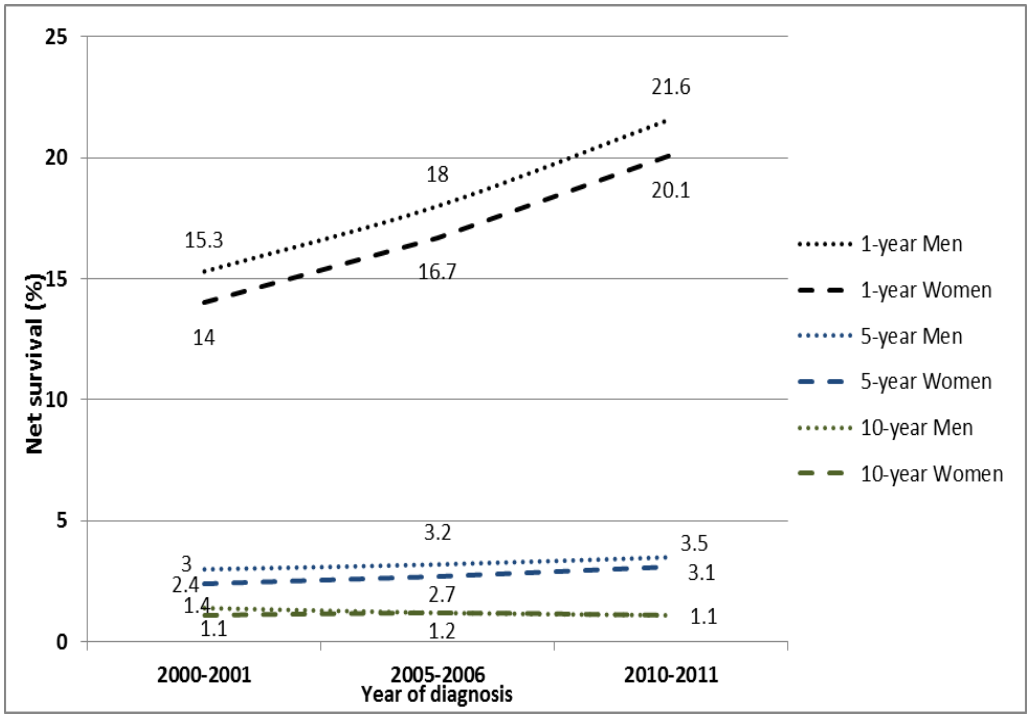
Note: Prepared by Cancer Research UK

Original data sources:

Survival estimates were provided on request by the Cancer Research UK Cancer Survival Group at the London School of Hygiene and Tropical Medicine. http://www.lshtm.ac.uk/eph/ncde/cancersurvival/

Net survival trends over time are plotted in Figure 4.3, from data provided by the London School of Hygiene and Tropical Medicine, summarised by Cancer Research UK. One-year, age-standardised net survival for pancreatic cancer increased from 15.3% in men and 14% in women in 2000–2001 to 21.6% and 20.1%, respectively, in 2010–2011. Longer-term survival figures are stable and poor; for example, 5-year survival in 2000–2001 was 3% in men and 2.4% in women, and 3.2% and 2.7%, respectively, in 2005–2006. Predicted 5-year survival for cancers diagnosed in 2010–2011 is 3.5% in men and 3.1% in women. Figures for 10-year survival suggest that approximately 1% of patients diagnosed with pancreatic cancer will survive their disease for at least 10 years.[53]

Figure 4.3 Pancreatic cancer age-standardised, net survival at 1 year, 5 years and 10 years, England and Wales, 2000–2011



Note:

Adapted from graphs prepared by Cancer Research UK

Original data sources:

Survival estimates were provided on request by the Cancer Research UK Cancer Survival Group at the London School of Hygiene and Tropical Medicine.
http://www.lshtm.ac.uk/eph/ncde/cancersurvival/

4.4.1.9 Conclusion

Pancreatic cancer is the tenth most common cancer in the UK, accounting for 3% of all new cases of cancer in 2012. In that year, it was the 13th most common cancer in men and the ninth most common cancer in women. The incidence of pancreatic cancer in the UK is rising, which may reflect the increased prevalence of obesity. Most patients with pancreatic cancer are diagnosed following emergency admission and have late-stage disease. This has consequences for both mortality and survival. Mortality is high, such that pancreatic cancer was the fifth most common cancer cause of death in the UK in 2012. Survival from pancreatic cancer remains poor – the 10-year net survival figure of 1% has remained unchanged since the 1970s.

4.4.2 Bladder cancer

4.4.2.1 The bladder

The urinary bladder is located in the pelvic cavity, immediately behind the pubic bones. Its function is to store urine that has been formed in the kidneys and, typically, it has a capacity of 500 ml. The urethra is far longer in men (20 cm)

than it is in women (just under 4 cm).[48] This is one of the main reasons why women are more susceptible than men to bladder infections.

4.4.2.2 Change in the definition of bladder cancer

Significantly, the definition of 'bladder cancer' changed in 1998, when the International Classification of Disease (ICD) for Oncology created separate diagnostic codes for carcinoma *in situ* of bladder (D09.0) and neoplasm of uncertain behaviour of bladder (D41.4), removing them from code C67 (bladder cancer).[62]

The change in definition has consequences for my study, as the bladder cancer cases were selected according to the pre-1998 definition of the disease. Subanalysis was conducted on cases that met the post-1998 definition of bladder cancer and their matched controls, and the risk estimates reported in Appendix 6.

The risk factors for bladder cancer (C67) are now discussed, as are its incidence, diagnosis, stage at diagnosis, mortality and survival.

4.4.2.3 Risk factors

Excluding genetic predisposition, there are a number of known external lifestyle factors that increase the risk of developing bladder cancer, the main ones in the UK being age (as reflected by the age-specific incidence rates discussed below) and tobacco smoke. In regions of the world where *Schistosoma haematobium* is

endemic, at least 41% of patients with bladder cancer have schistosomiasis; however, this infection is not a significant risk factor in the UK.[61]

According to government statistics, tobacco smoking is in decline. [63] The data were first collected in 1948 and showed that tobacco smoking was more prevalent among men (65%) than women (41%) (raw data not available). Both the overall prevalence of smoking and the gender gap have fallen steadily since then, with figures of 42% of men and 36% of women in 1980, and 31% of men and 28% of women in 1990. The latest data (2010) suggest that similar proportions of men (20%) and women (19%) smoke tobacco. This suggests that while smoking patterns will have contributed to historical gender-specific differences in the incidence of bladder cancer, this is likely to reduce in the future. Indeed, the percentage of incident cases of bladder cancer in the UK in 2010 attributable to tobacco smoking was estimated to be 37.5% in men and 34.4% in women.[50]

Other known risk factors for bladder cancer include occupational exposure, such as occurs in aluminium and in rubber production and in painting and decorating – industries traditionally dominated by men.[61,64]

4.4.2.4 Incidence

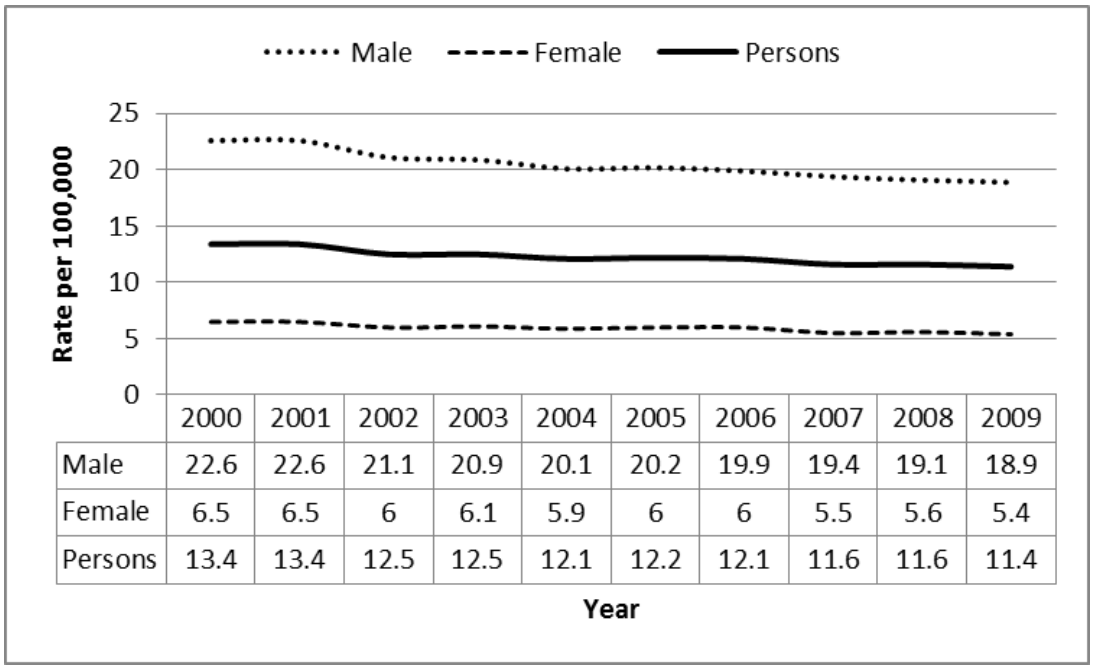
The latest statistics summarised by Cancer Research UK report that approximately 10,400 people (7,452 new cases in men; 2,947 in women) were diagnosed with bladder cancer (C67) in the UK in 2011. These figures equate to age-standardised incidence rates of 17.7 (95% CI: 17.3–18.1) per 100,000 men

and 5.4 (95% CI: 5.2–5.6) per 100,000 women. In terms of ranking, bladder cancer is the fourth most common cancer in men and the 13th most common cancer in women.[61]

Current incidence rates are low and similar in men and women below the age of 40 years. A gender difference in incidence develops as age increases above 50 years. After 60 years of age, bladder cancer incidence rises much more steeply in men than in women. Indeed, the *gender difference* in incidence rate peaks in the 65- to 69-year-old age group, at 70.5 per 100,000 men compared with 18.8 per 100,000 women. The absolute peak in age-specific incidence occurs in patients 85 years of age or older, at 264.3 per 100,000 men and 76.3 per 100,000 women.[61]

Aggregate incidence data for bladder cancer in each year of the recruitment period of my study are plotted in Figure 4.4. Cancer Research UK report that the incidence fell during this period, from 22.6 per 100,000 men and 6.5 per 100,000 women in the year 2000, to 18.9 per 100,000 men and 5.4 per 100,000 women in 2009. The decline in incidence over time is attributed to reductions in both exposure to occupational risk factors and tobacco smoking.[61]

Figure 4.4 European age-standardised incidence rates for bladder cancer (C67) per 100,000 population; UK data 2000–2009



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Male	22.6	22.6	21.1	20.9	20.1	20.2	19.9	19.4	19.1	18.9
Female	6.5	6.5	6	6.1	5.9	6	6	5.5	5.6	5.4
Persons	13.4	13.4	12.5	12.5	12.1	12.2	12.1	11.6	11.6	11.4

Note: Adapted from a graph prepared by Cancer Research UK. Original data sources:

1. Office for National Statistics. Cancer Statistics: Registrations Series MB1: http://www.ons.gov.uk/ons/search/index.html?newquery= series+mb1

2. Welsh Cancer Intelligence and Surveillance Unit: http://www.wcisu.wales.nhs.uk

3. Information Services Division Scotland. Cancer Information Programme: www.isdscotland.org/cancer

4. N. Ireland Cancer Registry: www.qub.ac.uk/nicr.

4.4.2.5 Diagnosis

There is no effective screening tool for bladder cancer, meaning that it is usually diagnosed when patients become symptomatic. In a recent meta-analysis of the association between symptoms and bladder or renal cancer in the primary care setting, Schmidt-Hansen *et al*[65] concluded that the only high-risk feature was visible haematuria. Five studies[6,66,67,68,69] were included in the meta-analysis, which reported a summary estimate of the positive predictive value of visible

haematuria for bladder or renal cancer of 5.1% (95% CI: 3.2–8.0%) in patients aged 15 to 100 years old. Drawing on this meta-analysis, The National Institute for Health and Care Excellence (NICE) recommend referral for investigation on a suspected cancer pathway for patients who are 45 years or older who have unexplained visible haematuria without a urinary tract infection, or visible haematuria that persists or recurs after successful treatment of a urinary tract infection.[9]

The National Cancer Intelligence Network[56] identified routes to diagnosis for the 42,924 bladder cancer diagnoses made in the period 2006 to 2010, i.e. under the previous[a] NICE referral guidelines.[57] The data indicate that 13,306 (31%, 95% CI: 31–32%) of the patients were diagnosed following referral via the two-week wait pathway. A further 12,349 patients (29%, 95% CI: 28–29%) were diagnosed following a routine or urgent referral made by the GP but not under the two-week wait route. Approximately one-fifth of patients (7,954/42,924= 19%, 95% CI: 18–19%) were diagnosed following emergency presentation.[56]

[a] Previous NICE guidelines were subtly different to the current ones, and recommended urgent referral for suspected urological cancer for patients of any age who presented with painless visible haematuria, once a urinary tract infection had been excluded. Urgent referral was also recommended for patients aged 40 years and older who presented with recurrent or persistent urinary tract infection accompanied by visible haematuria.

Cancer Research UK report a gender difference in emergency presentation, with a greater proportion of women (25%, raw data not reported) presenting via this route compared with men (16%, raw data not reported).[61]

Typically, bladder cancer is diagnosed in secondary care by flexible cystoscopy (see Section 4.5.1.6.2), with or without ultrasonography of the renal tract.[70]

4.4.2.6 Stage at diagnosis

The latest bladder cancer staging data summarised by Cancer Research UK relate to cancers diagnosed in England in 2013,[61] using data from the National Cancer Intelligence Network (see Table 4.3). Staging data were not complete, with the stage at diagnosis reported as 'unknown' for approximately one-quarter of all new diagnoses (2,228/8,775= 25.4%) in the analysis period. Over half (4,865/8,775= 55.4%) of all new bladder cancer cases were diagnosed at an early stage (i.e. stage I or II), leaving just under one-fifth being diagnosed at an advanced stage (i.e. stage III or IV) in England in 2013.

Table 4.3 Bladder cancer staging data, England 2013 (figures provided by the National Cancer Intelligence Network)

Stage at diagnosis	Number of new diagnoses (%) (*n* = 8,775 in total)
I	3,061 (34.9)
II	1,804 (20.6)
III	488 (5.6)
IV	1,194 (13.6)
Unknown	2,228 (25.4)
Total	8,775

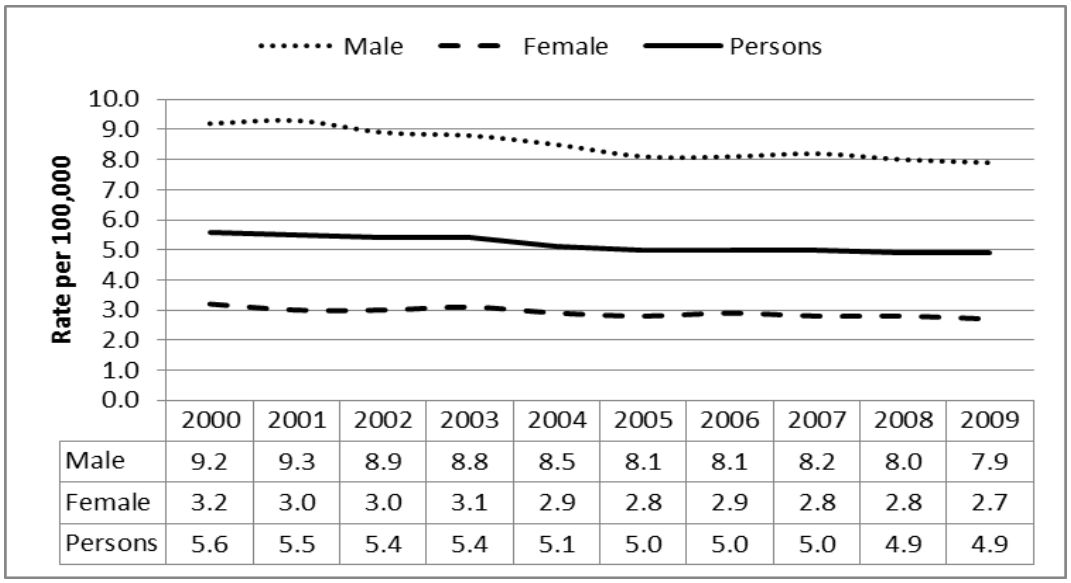
4.4.2.7 Mortality

The Office for National Statistics report that 4,536 deaths attributable to bladder cancer (C67) were registered in England and Wales in 2013, accounting for 3% of the 145,344 deaths from cancer registered that year. Gender-specific data indicate that roughly twice as many men (*n* = 3,039) as women (*n* = 1,497) died from bladder cancer in 2013, equating to approximately 4% of the 76,962 cancer deaths in men, and to 2% of the 68,382 cancer deaths in women. The majority (3,217/4,536= 70.9%) of all bladder cancer deaths occurred in patients who were 75 years of age or older.[60]

Bladder cancer was the 7th most common cancer cause of death in men, and the 13th most common in women in the UK in the previous year, 2012. Correspondingly, the European age-standardised mortality rate from bladder cancer in the UK in 2012 was 7.7 (95% CI 7.4–7.9) per 100,000 men and 2.8 (95% CI: 2.6–2.9) per 100,000 women.[61] Aggregate mortality data for bladder cancer in each year of the recruitment period of my study are plotted in Figure 4.5. The trend is downward, from initial values of 9.2 per 100,000 men and 3.2 per 100,000 women in 2000, to 7.9 per 100,000 men and 2.7 per 100,000 women in 2009. Indeed, Cancer Research UK report that age-standardised bladder cancer mortality rates fell in the UK between 2001–2003 and 2010–2012 by 15% in men and 11% in women. They attribute this to the decline in numbers of people who smoke tobacco, presumably reflecting the decline in bladder cancer incidence.[61] However, it is important to consider trends in survival before making such an interpretation, as trends in mortality only closely

mirror those in incidence when the majority of patients die shortly after diagnosis.[62]

Figure 4.5 European age-standardised mortality rates for bladder cancer (C67) per 100,000 population; UK data 2000–2009



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Male	9.2	9.3	8.9	8.8	8.5	8.1	8.1	8.2	8.0	7.9
Female	3.2	3.0	3.0	3.1	2.9	2.8	2.9	2.8	2.8	2.7
Persons	5.6	5.5	5.4	5.4	5.1	5.0	5.0	5.0	4.9	4.9

Note: Adapted from a graph prepared by Cancer Research UK

Original data sources:
1. Office for National Statistics, Mortality Statistics: Deaths registered in England and Wales: http://www.ons.gov.uk/ons/search/index.html?newquery= series+dr
2. General Register Office for Scotland, Deaths Time Series Data, Deaths in Scotland: http://www.gro-scotland.gov.uk/statistics/theme/vital-events/deaths/time-series.html
3. Northern Ireland Statistics and Research Agency, Deaths by cause: http://www.nisra.gov.uk/demography/default.asp14.htm

4.4.2.8 Survival

The 1-year, age-standardised, net survival rates for men and women diagnosed with bladder cancer (C67) in England and Wales during 2010–2011 are reported in Table 4.4. In this time period, survival at 1 year was considerably better in men (76.6%, 95% CI: 76.6–76.6%) than in women (61.6%, 95% CI:

61.5–62.8%). Predicted survival both at 5 and at 10 years is anticipated to be better in men than in women (Table 4.4).

Table 4.4 Bladder cancer (C67), age-standardised 1-, 5- and 10-year net survival, in adults aged 15–99 years, England and Wales, 2010-2011[61]

	Net survival (%, 95% CI) at		
	1 Year	**5 Years [a]**	**10 Years [a]**
Men	76.6 (76.6–76.6)	56.9 (56.7–57.1)	54.3 (53.9–54.8)
Women	61.6 (61.5–62.8)	45.6 (45.2–46.1)	39.5 (38.8–40.2)

[a] Predicted using an excess hazards model.

Note: Prepared by Cancer Research UK

Original data sources: Survival estimates were provided on request by the Cancer Research UK Cancer Survival Group at the London School of Hygiene and Tropical Medicine. ttp://www.lshtm.ac.uk/eph/ncde/cancersurvival/

The chances of surviving bladder cancer are strongly dependent on the stage at which it is diagnosed. Cancer Research UK provide relative survival data grouped by stage at diagnosis, estimated using data from The National Cancer Registration Service Eastern Office (formerly known as the Anglia Cancer Network). The data were collected during 2006 to 2010, coincident with much of my study period. Note, however, that these deaths relate to bladder cancer diagnoses that also included diagnostic codes D09.0 (carcinoma *in situ* of bladder) and D41.4 (neoplasm of uncertain behaviour of bladder). These cancers have a much better prognosis than cancers included in diagnostic code C67.

Relative survival was very high in those diagnosed with stage I disease, with 97% of men and 96% of women surviving for at least 1 year. Survival for at

least 5 years with a stage I diagnosis was also high, at 89% of men and 86% of women.

Relative survival decreased both in men and women the more advanced the stage at diagnosis, with the worst figures in those patients diagnosed at stage IV. Here, 26% and 9% of men reached 1-year and 5-year survival, respectively, compared with 25% and 11% of women. In general, relative survival was the same in men as in women, except at 1 year following a stage II diagnosis. Here, a higher percentage of men (74%) than women (57%) were alive at 1 year.

It is difficult to interpret the trends in bladder cancer survival over time, because of the above-described (see Section 4.4.2.2) change in definition of the disease in 1998. One-year age-standardised net survival data for bladder cancer suggest that survival has decreased since the 1990s; however, this is likely to be an artefact related to the exclusion of carcinoma *in situ* of bladder and neoplasm of uncertain behaviour of bladder, which are low-risk, benign tumours with a far better prognosis than cancers included in ICD10 code C67.

4.4.2.9 Conclusion

Bladder cancer is relatively common (it was the seventh most common cancer in the UK in 2011), but the incidence is higher in men than in women. The incidence has been falling both in men and women since the year 2000, some of which may be attributed to historical reductions in the prevalence of smoking. Most cases of bladder cancer are diagnosed at an early stage, and are associated with a high chance of survival. Despite this, bladder cancer was the

seventh most common cause of death in the UK in 2012, the majority of which occurred in patients aged 75 years or older. Mortality from bladder cancer is falling, partly a reflection of the reduced incidence. Trends in survival are difficult to interpret, because the definition of bladder cancer changed in 1998.

4.5 The symptoms

4.5.1 Haematuria

4.5.1.1 Formation of urine

Urine is formed by the kidneys and then transported via the ureter to the bladder, where it is stored prior to being expelled from the body during micturition. The first step in urine formation is the ultrafiltration of a plasma-like fluid from the blood into functional units called nephrons, of which the kidney contains approximately 1.2 million. In a healthy adult, the rate of ultrafiltrate formation is approximately 125 ml per minute or 180 litres per day, which equates to approximately 20% of renal blood flow. In healthy people, the ultrafiltrate from which urine is formed contains few red blood cells. The loss of around one million red blood cells in the urine daily, equating to about one cell per high-power field, is considered normal.[71]

4.5.1.2 Definition

Haematuria is defined as the abnormal presence of red blood cells in urine, either in sufficient quantities to be readily recognised by the patient (**visible**) or in an amount so small that it requires detection by urine dipstick testing or

microscopy (**non-visible**).[72] Visible and non-visible are the terms currently recommended to describe the two types of haematuria and so have been adopted consistently throughout this thesis in preference to other descriptors such as ‘frank’, ‘gross’ or ‘macroscopic’ (visible) and ‘invisible’, ‘microscopic’ or ‘dipstick-positive’ (non-visible).[73]

The threshold for definition of non-visible haematuria is three or more red blood cells in urine per high-power field on microscopy, in the absence of infection or proteinuria.[72,74] Non-visible haematuria is described as symptomatic when it is accompanied by lower urinary tract symptoms such as hesitancy, frequency, urgency or dysuria, whereas it is called asymptomatic when discovered incidentally in the absence of any accompanying symptoms.[73]

4.5.1.3 Background rate

The background rate of visible haematuria in the general population is unclear. Summerton estimated that a UK GP will encounter less than one case per year, but it is not clear how this figure was calculated. He also quoted a figure of 0.6 cases per year per GP in a standard practice in the Netherlands, according to morbidity figures from general practice.[75]

There is also a great deal of uncertainty about the prevalence of non-visible haematuria in the general population.[76] This may relate to the small number of studies conducted and heterogeneity in both the method of detection (dipstick test alone or in combination with microscopy) and the number of tests carried out to confirm the diagnosis.[77] Four screening studies are discussed below;

however, it is important to note that any inferences about the prevalence of non-visible haematuria in the general population should be drawn with caution, because all these studies used non-probability sampling methods.

In a sample of 20,751 men and women in the USA undergoing dipstick urinalysis screening as part of a private health assessment, the overall frequency of asymptomatic non-visible haematuria was 2.9%. The study investigated the relationship of non-visible haematuria with gender and age: it tended to decrease with age, being 5.0% in women aged 55–75 years, falling to 3.4% in women >75 compared with respective values of 1.8% and 1.08% in men. Non-visible haematuria also tended to be more prevalent in women than men overall. It should be noted, however, that the sample included different age ranges for the two genders (men ≥35 years; women ≥55 years), women aged <55 years being excluded owing to problems associated with misattribution of vaginal bleeding.[78]

In a UK retrospective study of 10,050 men attending a private clinic for health screening between February 1983 and January 1984, the prevalence of non-visible haematuria was 2.5%. Haematuria was detected in freshly collected midstream urine sample using dipsticks sensitive to haemoglobin, and confirmed by re-testing and microscopy within 4 hours. The age range of men discovered to have haematuria was 21–72 years (summary statistics not reported).[79]

In a prospective study, 855 elderly men (aged 60–85 years) were invited to attend an inner city health centre in Leeds, UK, for urine screening as part of

general health check. In total, 578 men responded, of whom 78 (13%) had dipstick-positive haematuria on the initial test conducted by a nursing sister. All participants were asked to test their urine weekly for 10 weeks, which resulted in a further 54 men (9%) reporting a positive result. The authors combined these results to report an overall prevalence of 132/578, or 23%, in men >60 years.[80]

Finally, a retrospective screening study examined 1000 healthy Israeli military male recruits, aged 18–33 years at the beginning of a 15-year investigation.[76] Non-visible haematuria was defined as two to four or more red blood cells per high-powered field on microscopical examination of urine. The point prevalence of non-visible haematuria was 5.2% (results from 12,227 urinalyses conducted over the 15-year study). The study reported a cumulative incidence of non-visible haematuria of 387 in the 1000 men (38.7%) on one or more occasion, and of 161/1000 (16.1%) on two or more occasions. The authors acknowledged the limitation that the sampling of young adult men from the military prevents inferences to be made about the general population. This value may overestimate the frequency of non-visible haematuria in the general population, which tends to be more sedentary than military recruits, given the known association between non-visible haematuria and vigorous exercise.[81]

4.5.1.4 Detection

Detection of haematuria relies on both observation and urinalysis, including visual inspection, chemical dipstick testing and microscopy.

4.5.1.4.1 Visual inspection of urine

Visible haematuria is an alarming symptom, and presents as smoky, bright red or reddish-brown urine, which may or may not contain blood clots.[82] The overt nature of visible haematuria means that it has a very low false-positive rate of detection. However, patients may mistake other conditions as being haematuria, such as rectal or vaginal bleeding, haemoglobinuria (excretion of free haemoglobin in the urine), myoglobinuria (urinary excretion of myoglobin from broken down muscle cells) and acute intermittent porphyria (fresh urine appears normal, but develops a dark-red colour if left to stand). Finally, red/pink or orange coloration of urine may occur after ingesting beetroot, blackberries and some medications (e.g. rifampicin).[82] Misattribution as haematuria can be avoided by careful dietary and medication history-taking, as well as by a selection of tests (Table 4.5).[49]

Table 4.5 Tests to distinguish true visible haematuria from conditions that mimic it[49,82]

Condition	Test to distinguish it from visible haematuria
Haemoglobinuria	Microscopy of a fresh urine sample – red blood cells are not present in the urine
Myoglobinuria	Microscopy of a fresh urine sample – red blood cells not present in the urine Specific laboratory tests are available to distinguish myoglobin from haemoglobin
Acute intermittent porphyria	Examine fresh urine samples Red blood cells not detected in urine

Non-visible haematuria is not detectable on visual inspection, but relies on examination of urine using chemical dipsticks sensitive to haem or, less frequently, by urine microscopy.

4.5.1.4.2 Chemical dipstick testing

Dipstick tests detect the presence of haem, which is a prosthetic group that forms the non-protein part of haemoglobin and myoglobin, enabling both to bind oxygen.[83] Therefore, chemical dipsticks detect intact red blood cells (shown as green spots on the reagent strip), haemoglobin (green-coloured fields) or free myoglobin (also as green-coloured fields).[82] They are calibrated to give a negative finding at the normal level of blood loss in urine. Several chemical dipsticks are available commercially, and all give a result immediately in GP surgeries.

There are few reliable data available for calculating dipstick performance characteristics. Sensitivity is reported to range between 91% and 100%, but with a specificity ranging from 65% to 99% for two to five red blood cells per high-power field, the potential for false-positives may be as great as 35%.[84] Causes of false-positive readings for haematuria include contamination with menstrual blood or with bleach, patient dehydration (conserving water inevitably increases the concentration but not the overall number of red blood cells), myoglobinuria (e.g. following breakdown of muscle tissue, such as can occur in patients taking statins or recreational drugs, or following trauma) and using a

stale urine sample.[82,85] Inappropriate storage of dipsticks increases the likelihood of false-negative readings, as does an acidic urine and bacteriuria.[85]

Table 4.6 Indications for chemical dipstick testing [82]

Type of use	Indication	Of value in
Screening	Random	Diabetes mellitus
		Asymptomatic bacteriuria
	Selective	Antenatal care
		Hypertension
Diagnosis	Primary renal disease	Glomerulonephritis
	Secondary renal disease	Bacterial endocarditis
	Non-renal disorders	Diabetes mellitus
Monitoring	Disease progression	Diabetic nephropathy
	Drug toxicity	Gold therapy
	Drug compliance	Rifampicin therapy
	Illicit drug use	Opioids, benzodiazepines

This table was reproduced, with permission, from *Macleod's Clinical Examination*, 12th edn., Douglas, G., Nicol, F., Robertson, C. (eds.), pp.1–476, Copyright Elsevier (2009).

Despite the potential for inaccuracy, a recent systematic review concluded that dipsticks are a reasonable method to use in isolation to detect non-visible haematuria. Data from 17 studies evaluating the accuracy of dipstick tests in the diagnosis of haematuria, compared with the reference standard of microscopy, were pooled. The positive likelihood ratio of dipsticks to diagnose haematuria was estimated to be 5.99 (95% CI: 4.04–8.89) and the negative likelihood ratio

0.21 (95% CI: 0.17–0.26), with the proviso that the figures should be interpreted cautiously, owing to significant heterogeneity between studies.[72] Clinically, dipstick testing is widely indicated in screening, diagnosis and monitoring of disease (Table 4.6).[82]

4.5.1.4.3 Urine microscopy

On visual inspection, the colour of urine varies with the patient's hydration status – the lighter the colour, the more hydrated the patient. Regardless of its colour, normal fresh urine is clear, but cloudiness may develop in a sample that has been left to stand for long enough that phosphate and urate salts precipitate out of solution. For this and other reasons, it is important to conduct urinalysis on fresh samples. Cloudiness of a freshly collected urine sample is abnormal, and indicates the presence of pus and/or bacteria.[82]

Early morning, midstream urine samples need careful preparation before examination. Best practice is as follows: centrifuge a 10-ml sample at 2,000 rpm for 5 min, discard the supernatant and re-suspend the sediment in 0.5–1.0 ml of the original urine sample. Examine a single drop of the freshly prepared sample under a microscope.[84,86] Urine microscopy detects non-visible haematuria far more accurately than chemical dipsticks, with a very low false-positive rate, but only when performed by trained technicians or nephrologists as described above.[73] It is utterly impractical to provide this level of service in primary care, and so urine microscopy is rarely carried out in this setting.

4.5.1.5 Causes

Blood in urine can originate from anywhere along the urinary tract, but sources are generally grouped according to the potential underlying pathology as being glomerular (nephrological) or non-glomerular (urological).[72]

Visible haematuria usually has a non-glomerular source of blood, such as the ureter, bladder or urethra. Visible haematuria associated with loin pain (renal colic) is often caused by a kidney stone as it is passed from the body in the urine. Blood that clears before the urine is fully voided usually indicates an origin in the urethra, whereas an association with frequency or dysuria commonly indicates the bladder as being the source.[82] Tumours themselves are liable to bleed because they have a fragile blood supply. When they become large enough to invade surrounding tissue, tumours may damage healthy vasculature and also result in bleeding. For tumours sited anywhere along the urinary tract, the blood is expelled in the urine and is apparent to the patient as visible haematuria.

Causes of non-visible haematuria, without proteinuria, can be either glomerular (nephrological) or non-glomerular (urological) (see Table 4.7).

Table 4.7 Causes of isolated non-visible haematuria (causes are listed in order of descending frequency of presentation, according to available data)

Origin	Age	
	< 50 years	**≥50 years**
Glomerular	IgA nephropathy	IgA nephropathy
	Thin basement membrane disease (also known as benign familiar haematuria)	Hereditary nephritis (also known as Alport syndrome)
	Hereditary nephritis (also known as Alport syndrome)	Mild focal glomerulonephritis of other causes
	Mild focal glomerulonephritis of other causes	
Non-glomerular		
Upper urinary tract	Nephrolithiasis	Nephrolithiasis
	Pyelonephritis	Renal-cell cancer
	Polycystic kidney disease	Polycystic kidney disease
	Medullary sponge kidney	Pyelonephritis
	Hypercalciuria, hyperuricosuria, or both, without documented stones	Renal-pelvis or ureteral transitional-cell cancer
	Renal trauma	Papillary necrosis
	Papillary necrosis	Renal infarction
	Ureteral stricture and hydronephrosis	Ureteral stricture and hydronephrosis
	Sickle cell trait or disease in blacks	Renal tuberculosis
	Renal infarction or arteriovenous malformation	
	Renal tuberculosis in endemic areas or in patients with HIV infection	

Origin	Age	
	< 50 years	**≥50 years**
Lower urinary tract	Cystitis, prostatitis and urethritis	Cystitis, prostatitis and urethritis
	Benign bladder and ureteral polyps and tumours	Bladder cancer
	Bladder cancer	Prostate cancer
	Prostate cancer	Benign bladder and ureteral polyps and tumours
	Urethral and meatal strictures	
	Schistosoma haematobium in North Africans	
Uncertain	Exercise	Exercise
	'Benign' (unexplained)	Over-anticoagulation (usually with warfarin)
	Over-anticoagulation (usually with warfarin)	
	Factitious haematuria (usually presents with visible haematuria)	

HIV, human immunodeficiency virus.

Reproduced with permission from Cohen and Brown (2003), Copyright Massachusetts Medical Society.[77]

Causes of haematuria are classified as renal (e.g. neoplasia, glomerulonephritis, tubulointerstitial nephritis, polycystic kidney disease, papillary necrosis, infection and trauma) or extrarenal (e.g. calculi, infection, neoplasia and trauma), each group consisting of a mix of benign and serious conditions.[85] Infection and calculi are the most common non-malignant causes in the over 40s – the patient group included in our study. Urinary tract infection

(UTI) is very common – more so in women than men owing to their short urethra.

Completing menopause further increases the risk in older women, as loss of oestrogen renders the urinary tract more vulnerable to infection. Calculi are less common than UTI (lifetime incidence up to 15%) and presentation peaks in a younger age group (20–40 years) than is included in my study.[85] Therefore, the most common benign cause of haematuria in the over 40s is UTI. In addition, in women, vaginal bleeding may be mistaken for haematuria.

Serious pathology resulting in haematuria includes neoplasms such as bladder cancer. Bladder cancer is more common in men than in women (see Section 4.4.2.4).[34] The most highly predictive symptom of bladder cancer is visible haematuria.[4,6,67,69] This is reflected by current UK referral guidelines published in 2015 by the National Institute of Health and Care Excellence (NICE), which recommend referral for patients over 45 years of age who have unexplained visible haematuria without UTI, or visible haematuria that persists or recurs after successful treatment of UTI.[9]

4.5.1.6 Investigation

A systematic review conducted in 2006 to determine an effective diagnostic strategy for the investigation of visible and non-visible haematuria concluded that there were ‘insufficient data currently available to derive an evidence-based algorithm’. The authors presented a diagnostic algorithm based on the opinion and practice of clinical experts.[72] Urological referral and investigation for

urological cancer using ultrasound (US), cystoscopy and cytology was recommended for:

- All patients with visible haematuria
- Patients with asymptomatic non-visible haematuria accompanied by risk factors for cancer (such as smoking)
- All patients with persistent symptomatic non-visible haematuria

4.5.1.6.1 Ultrasound

Ultrasound (US) is a non-invasive imaging technique that is used in the investigation of haematuria. It is excellent at identifying renal cysts but has limited usefulness in detecting solid renal masses that are smaller than 3 cm in size.[87] Its performance is highly dependent on the expertise of the ultrasonographer,[72] and its sensitivity at detecting bladder cancer in patients was reported to be 19%.[87,88]

4.5.1.6.2 Cystoscopy

Cystoscopy is an invasive investigation carried out by urologists in secondary care. It is indicated when a non-glomerular cause of haematuria is suspected, to identify malignancy – particularly bladder cancer.[87] The sensitivity of cystoscopy to detect bladder cancer, as indicated by haematuria alone, was reported to be 95.6% (95% CI: 87.2–98.6%), which equates to a positive predictive value of haematuria for bladder cancer of 65.0% (95% CI: 55.2–73.7%). The specificity

was 94.3% (95% CI: 92.1–95.9%), which is reassuringly high for an investigation used to rule out malignancy.[89]

4.5.1.6.3 Cytology

Cells sloughed from urothelial tumours are excreted in the urine and can be detected by cytology. However, the reliability of this technique depends greatly on the technical proficiency of the pathologist examining the samples. The sensitivity of cytology to detect bladder cancer ranges from 40% to 70%; therefore, while a positive result is considered diagnostic of the presence of urothelial cancer, a negative result cannot rule out a malignancy, owing to the test's high false-negative rate.[87]

4.5.1.7 Conclusion

There are two forms of haematuria: visible and non-visible. Visible haematuria may have a benign or a malignant underlying pathology. However, for patients and GPs alike, it is an alarming symptom not least because of its strong association with urological tumours. Non-visible haematuria is only apparent on testing of urine, and the majority of cases have an underlying benign pathology. Neither form of haematuria has an association with pancreatic cancer.

4.5.2 Jaundice

4.5.2.1 Introduction

Jaundice, also known as icterus, occurs when plasma levels of bilirubin, normally 3–17 µmol/l, rise in excess of 40 µmol/l.[90,91] Bilirubin is a by-product of

the breakdown of red blood cells, is yellow in colour, and is excreted in both bile and urine.[83]

4.5.2.2 Detection

Moderately raised levels of plasma bilirubin require detection by liver function blood tests. As its plasma levels rise in excess of 40 µmol/l, the yellow-coloured bilirubin becomes noticeably visible in the whites of the eyes, the mucous membranes and eventually the skin and urine.[90]

4.5.2.3 Background rate

In a cohort study of 186,814 adults aged over 45 years, Taylor *et al*[92] reported an annual incidence of jaundice of 0.74 per 1,000 patients. An acknowledged limitation of the Taylor study is that their cohort was not a random selection of all adults aged over 45 years, as it over-represented the age group who suffer cancer. This is because it was carried out as part of the Discovery Programme,[a] and the cohort was selected from the group of control patients from the CAPER studies, who were matched on age, sex and GP practice to cancer cases.[92]

4.5.2.4 Causes

There are three main mechanisms by which hyperbilirubinaemia can lead to jaundice. First, excess production of bilirubin at a rate that overwhelms the

[a] As described in Section 4.2, the Discovery Programme used CPRD data in case–control studies to characterise the presentation of 13 common cancers in primary care in the UK.

liver's ability to process it for excretion (termed pre-hepatic jaundice). Secondly, liver malfunction (hepatic jaundice), and, finally, obstruction of the bile duct, one of the main routes of bilirubin excretion (post-hepatic jaundice).[91]

4.5.2.4.1 Causes of pre-hepatic jaundice

The most common cause of pre-hepatic jaundice is excessive breakdown of red blood cells, such as occurs in haemolytic anaemia, Gilbert syndrome, spherocytosis, sickle cell disease and thalassaemia major.[90]

4.5.2.4.2 Causes of hepatic jaundice

Common forms of hepatic jaundice include viral hepatitis and alcoholic hepatitis, and it may also be associated with alcoholic cirrhosis, primary biliary cirrhosis and the ingestion of certain drugs.[90]

4.5.2.4.3 Causes of post-hepatic jaundice

Post-hepatic jaundice commonly arises when the common bile duct becomes obstructed. The obstruction is commonly caused by a benign gallstone; however, it may be caused by an impinging malignancy of pancreatic origin.[90,93]

4.5.2.5 Investigations

In brief, initial investigations in suspected jaundice include urine and blood tests to determine whether the patient has hyperbilirubinaemia and to assess liver function. Investigations escalate to include abdominal imaging using a variety of modalities to determine whether the bile duct is obstructed.[93]

4.5.2.6 Conclusion

Jaundice has a variety of causes in adults, ranging from benign conditions such as gallstones to life-threatening malignancies such as pancreatic cancer. As it may indicate the presence of a serious condition, it is viewed as a worrying symptom in adults, particularly in those confirmed as not having gallstones or in those with other features suggestive of pancreatic cancer.[93] It has no known association with bladder cancer.

4.5.3 Abdominal pain

4.5.3.1 Introduction

Pain has been described as ‘an unpleasant sensory and emotional experience resulting from a stimulus that has already caused, is causing, or is likely to cause tissue damage’.[94] The abdomen is defined as the region of the trunk between the diaphragm and the inlet of the pelvis. It contains vital organs (viscera, such as the gastrointestinal tract, the liver, biliary ducts and pancreas), parts of the urinary system, as well the aorta and its branches, the inferior vena cava and its tributaries and the portal vein.[48] Therefore, it was reasonable to assume that both pancreatic and bladder cancer may present with abdominal pain in the design of the studies.

In a clinical examination, pain is characterised by its severity, nature and location, each of which is discussed below in more detail.

4.5.3.2 Assessment of pain severity

Assessment of pain severity relies on how it is perceived and reported, both of which depend on the patient's subjective and emotional responses to it. Indeed, pain severity is such an individual experience that some regard attempts to categorise it as unhelpful and counterproductive.[94,95] Despite this, various methods have been developed to measure pain severity, including scales that require patients to match it to a number, to a visual image or to a verbal description. For chronic pain, methods must be extended to include a measure of the pain's impact on the patient's physical and emotional well-being, as well as their social functioning.[96] There are no Read codes for severity of pain, although there are codes to indicate when a patient has completed a McGill, Oswestry or Dallas pain scale, a visual analogue pain scale or a pain diary. Indeed, a retrospective, cohort study of the quality of medical record keeping conducted in 18 general practices in Exeter, UK, reported that the recording of symptom severity is poor in computerised systems. Of 2,444 individual symptom codes documented in electronic medical records, only 290 (11.9%, 95% CI: 10.6–13.2%) included an indication of its severity.[11]

Given the highly individual and subjective nature of pain perception, and the poor documentation of its severity, information about pain severity as a marker of bladder or pancreatic cancer was not sought.

4.5.3.3 The nature of abdominal pain

In broad terms, the nature of abdominal pain is determined by whether it arises from inflammation or obstruction. Abdominal pain associated with inflammation tends to be constant, of varying severity and is exacerbated by physical disturbance. In contrast, pain arising from obstruction of a muscular tube (such as the bowel or ureter) is ‘colicky’, i.e. it fluctuates in severity and comes in waves, and is described as ‘gripping’. Prolonged obstruction can lead to distension of the bowel or ureter, at which point the pain is described not as colicky, but as constant and ‘stretching’.[91]

4.5.3.4 The significance of pain location

4.5.3.4.1 Introduction

The abdominal organs are closely positioned within the abdominal cavity; therefore, disease in one organ readily affects the others. In addition, many of the abdominal organs are inaccessible to palpation, because either they lie deep within the abdominal cavity, or are afforded bony protection by the ribs, pelvis or spine. Despite these limitations, vital diagnostic information can be obtained from a thorough clinical examination and history.[91]

4.5.3.4.2 Anatomy

Anatomically, the abdomen is divided into three zones horizontally (upper, central and lower) and three vertically (right, central and left), as depicted schematically Figure 4.6.

Figure 4.6 Schematic diagram to show the regions of the abdomen and the zones where pain from the pancreas and bladder (underlined) is experienced

Right hypochondrium Gall bladder	**Epigastrium** Stomach & duodenum <u>Pancreas</u>	**Left hypochondrium** <u>Pancreas</u>
Right lumbar Kidney	**Umbilical** **O** Small bowel, caecum, retroperitoneal structures	**Left lumbar** Kidney
Right iliac fossa Appendix & caecum	**Hypogastrium (aka suprapubic region)** Transverse colon, <u>bladder</u>, uterus & adnexae	**Left iliac fossa** Sigmoid colon

A common alternative to the terms described above is to divide the abdomen into four quadrants – upper and lower, left and right – demarcated by the midline vertically and the umbilicus horizontally. GPs may use any of these terms to describe the location of abdominal pain, a fact that needs to be borne in mind when searching for abdominal pain records.

4.5.3.4.3 Visceral pain

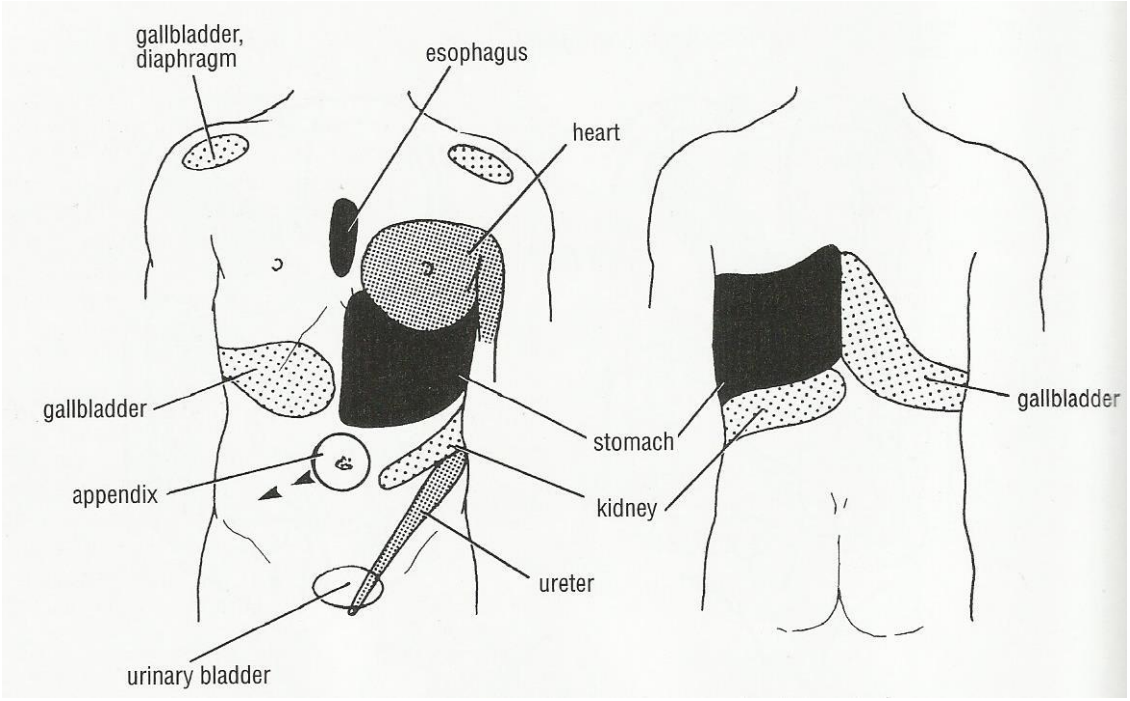
While much visceral pain is simply felt ‘centrally’, the pain arising from some visceral structures is perceived in characteristic locations; for example, pain from the pancreas is felt in the left hypochondrium and centrally in the epigastrium, whereas a painful bladder is located in the hypogastrium.

Pain may radiate away from these specific sites in characteristic ways as surrounding structures become affected by the disease; for example, pain originating in the pancreas radiates through to the back and to the left. Therefore, patterns of pain radiation are helpful for identifying not just the source of pain but also the extent of underlying disease.[91]

4.5.3.4.4 Referred abdominal pain

Pain from abdominal structures may also be referred to other parts of the body that share the same nerve supply, as shown in Figure 4.7. Referral of pain in this way derives from neural connections that develop in the embryo, good knowledge of which permits clinicians to identify where the underlying disease is likely to be.[48,91]

Figure 4.7 Some important skin areas involved in referred visceral pain (Figure 5-70 reproduced with permission from Richard S. Snell, Clinical Anatomy for Medical Students, 5th edition, © Richard S. Snell MD PhD, 1995[48])



For example, visceral pain from the gallbladder is not just perceived in the right hypochondrium and radiating right and through to the back – it also may be referred up to the right shoulder tip, where it is perceived by the patient as an entirely dissociated pain.[48]

Consequently, the signature locations of abdominal pain associated with pancreatic cancer generally differ from those that might be experienced by a patient with bladder cancer. This was important when deciding which pain codes GPs might choose to record abdominal pain, and therefore which pain codes to select for study (see Section 6.8.1).

4.5.3.4.5 Pain from abdominal wall structures (somatic pain)

Pain arising in abdominal wall structures is described as ‘somatic’. Relevant structures include the skin, fascia, muscles and, importantly, the outer layer of the membrane lining the abdomen – known as the parietal peritoneum. The nerves transmitting pain from the parietal peritoneum also innervate the overlying skin and mediate local reflexes. Therefore, abdominal pain arising from an inflamed parietal peritoneum may present with hyperaesthesia (excessive physical sensitivity), tenderness of the skin and increased abdominal tone, called ‘guarding’. It may also be characterised by tenderness experienced upon sudden release of an abdominal wall stretch induced by deep palpation – so-called rebound tenderness.[48]

Bladder and pancreatic cancer are not reported as being associated with somatic abdominal pain.

4.5.3.5 Investigation for abdominal pain

In cancer diagnosis, abdominal pain is a ‘low-risk but not no-risk’ symptom. It has a high enough positive predictive value for colorectal, oesophageal, stomach, ovarian or pancreatic cancer to warrant either investigation in primary care or referral. For suspected pancreatic cancer, the advice relating to patients aged 60 years and older, who have abdominal pain in conjunction with weight loss, is to consider an urgent computed tomography scan – to be performed within 2 weeks – or an urgent ultrasound scan if the former is not available.[9]

4.5.3.6 Conclusion

Determining the location, nature and patterns of radiation of abdominal pain is helpful for identifying the source of pain and the extent and nature of underlying disease. While abdominal pain is a feature of both bladder and pancreatic cancer, its characteristics vary with the cancer site.

5 The research questions

To round out the introductory part of this thesis, I thought it would be helpful to draw together the various strands that have informed the research questions that this thesis is designed to address.

First, there is existing evidence to suggest that GPs use text fields to record a lot of information about symptoms, and that this is hidden to the majority of CPRD-based studies because they restrict their analysis to codes. Therefore, at the very least, CPRD studies are likely to under-report symptom prevalence. Secondly, little is known about GPs' choice of recording style and whether this varies with the type of patient, the type of symptom or the context in which it presents – all of which have the potential to introduce bias. Thirdly, policy decisions that decide the allocation of public money to healthcare provision are based on evidence provided by primary care research, an increasing amount of which is conducted using CPRD data. Therefore, such decisions need to be made in full knowledge of any under-reporting or bias introduced by the omission of text data. With this in mind, the following research questions were formulated.

1. **How much symptom information is documented in electronic medical records using text rather than a code?**

Most CPRD practices use ViSion (ViSion INPS, London, UK) to record the electronic medical record, in which GPs must choose a Read code first, after which a comments box opens. Here, GPs can type freely and are not limited to

elaborating on the code. From clinical and medico-legal standpoints, codes and text are equally accessible. The same cannot be said for research – while codes are fully and routinely available to researchers, text records are not. Therefore, researchers are oblivious to anything that is recorded *only* as an inaccessible comment (henceforth called 'text-only records').

2. **Are studies of coded data vulnerable to bias arising from the differential use of text and codes between comparison groups?**

Systematic differences in the use of text rather than code to record clinical events between comparison groups will lead to bias in studies restricting analysis to codes, because of unequal data collection. For example, if text-only recording is greater for controls than cases, outcome measures will be artificially inflated because clinical events experienced by the control group will be consistently under-reported. Conversely, outcome measures will be artificially reduced if text-only recording is greater for cases than controls.

3. **Does recording style vary with type of symptom?**

Some symptoms are known 'alarms' for disease, whereas other symptoms may not trigger such a high level of clinical suspicion. To the best of my knowledge the effect of a symptom's clinical significance on GPs' choice of recording style (codes or text) has not been quantified.[13,15]

Presenting with an alarm symptom increases the likelihood that a GP will instigate an investigation, carry out some tests or perhaps make a referral. Such actions tend to require GPs to use codes in their record-keeping.[3,97] Therefore,

it is feasible that GPs are more likely to use codes for alarm symptoms than they are for symptoms perceived to be of little clinical significance. If this were the case, studies conducted using coded data only would be biased towards alarm symptoms and, conversely, against relatively innocuous ones.

4. **Does the recording style vary with the clinical context of presentation of a symptom?**

During a consultation GPs draw on their knowledge and experience to formulate a working diagnosis that best fits the patient's presentation, taking into account patient characteristics such as their gender, age and lifestyle. The dual-process model of diagnostic reasoning has been proposed relatively recently, involving a mixture of intuitive and analytical thought processes.[98] It may be hypothesised that this process leads doctors to record preferentially any clinical features substantiating what they think is the correct clinical diagnosis, rather than to record absolutely everything reported by the patient. This may muddy the interpretation of how alarm symptoms are treated – after all, a symptom may be considered a red flag when presented with a constellation of other symptoms that 'fit' the working diagnosis, but a red herring when it is incongruous. Such recording behaviour would lead to bias in the capture of information about symptoms based on their perceived likelihood of presentation in a particular context.

5. **Do the text data provide additional value to coded data?**

Research questions 1–4 address the issue of bias introduced by ignoring text data within the CPRD. However, it is also important to investigate whether there is a difference in the ‘research quality’ of code- and text-based variables. For example, do code- and text-based variables give similar measures of association between symptoms and cancer in regression analysis?

6 Methods

These case–control studies extended two recent CAPER studies of CPRD data that reported the main features of bladder[4] and pancreatic[5] cancer presentation in primary care. The extension was in the form of supplementation with text records for three features – haematuria, abdominal pain and jaundice – to identify what information had been missed because it had not been coded. This required cross-checking of dates to establish whether the text note was elaborating on an event already recorded by the GP using a Read code or indeed was the sole documentation of the symptom. This date matching could not be done if the new text-based variables were simply bolted on to existing datasets[4,5]; instead, the datasets had to be re-created using the same study participants.

It was decided to extend the original studies conducted in the CPRD rather than start afresh with a completely new dataset or change to a different dataset (e.g. QResearch or THIN, see Section 2.6). This is because the original studies, along with others also using CPRD data collected between 1 January 2000 and 31 December 2009 under the same design,[41,42,43,44] were used as evidence in 2015 revision of the National Institute of Health and Care Excellence referral guidelines for suspected cancer.[9] This enabled my study to identify and quantify bias arising from data loss in the original studies and to revise the risk estimates for bladder and pancreatic cancer accordingly.

The methods specific to this study, including pre-analysis work-up, dataset re-creation and identification of text-only recording, are detailed below. For completeness, the setting, methods of recruitment, and design of the original studies are also described.

6.1 The setting

Participants for both studies were selected from the CPRD at a time when it was collecting data on approximately 12 million patients across the UK, equating to 8% of the population. The CPRD's probability sampling methods and geographical representativeness allow inferences about the UK population as a whole to be drawn from these studies' results.

6.2 The recruitment period

The recruitment period was 1 January, 2000 to 31 December, 2009 inclusive for both studies. The end point was determined as the last date for which the Office for National Statistics for England and Wales had complete data for the numbers of cases of bladder and pancreatic cancer at the time the original studies were devised.[4,5] Ten years was considered an adequate length of time to study the two cancers.

This recruitment period also coincided with increased automation of test result transmission from the laboratory to the GP surgery. Laboratory test results used to be transferred manually into the patient record, but 'drag-and-drop' from an

email was introduced in 2003. However, since 2007 data entry of all laboratory results has been fully automated, linked through the patient's NHS number.[99]

6.3 Study participants

Participants were eligible for potential inclusion in the studies if they were registered during the recruitment period with a CPRD practice whose data met 'up-to-date standard' quality criteria (see Section 2.4.4). As the CPRD is a probability sample of the UK population (see Section 2.4.2), it is fair to assume that the cases are representative of typical cancer patients. In addition, cases and controls are all drawn from the same underlying population. For these reasons, the study design minimises, if not eliminates, selection bias.

6.3.1 Case ascertainment

Potential cases were identified as those patients with a clinical or referral record of incident bladder or pancreatic cancer in the recruitment period. Incident cancer was defined by the Read code list provided by the primary investigator (Professor Willie Hamilton) and agreed with the CPRD research team (complete lists in Appendix 15.2).

The definition of bladder cancer reflected that used *before* the change in disease definition introduced by International Classification of Disease (ICD) for Oncology in 1998, i.e. it included carcinoma *in situ* of bladder and neoplasm of uncertain behaviour of bladder (see Section 4.4.2.2).

Read codes are mapped 1:1 to medcodes by the CPRD purely to facilitate analysis in Stata and essentially the terms are interchangeable.[a] The list was used to create a 'disease thesaurus' of medcodes in Stata, in which each cancer site was assigned a unique number. Medcodes pertaining to the diagnosis of a specific cancer could therefore be extracted by searching on the cancer site code.

6.3.2 Inclusion criteria for cases

Inclusion criteria for cases consisted of:

- Being aged ≥40 years at diagnosis
- Being registered at the practice for at least 1 year before diagnosis
- Having gender clearly recorded as either 'male' or 'female'

6.3.3 Exclusion criteria for cases

Exclusion criteria for cases consisted of:

- Having a secondary cancer
- Being diagnosed with any cancer before 1 January, 2000

[a] In this thesis, I use Read codes in the context of the GP consultation, as these are what the GPs use to create the medical record, whereas I use medcode in the context of research; however, I should emphasise that the two terms are completely interchangeable.

- Not consulting in primary care in the year before the cancer diagnosis
- Not having any matched controls

6.3.4 Final selection of cases

Inclusion and exclusion criteria were applied, after which a maximum of 5,000 cases with the relevant cancer diagnosis were recruited to the bladder cancer study using random sampling. The limit of 5,000 cases was set, in order to keep within the financial resources of the original studies.[4,5] Fewer than 5,000 patients received a diagnosis of pancreatic cancer; therefore, all those who both met the inclusion criteria and were not ruled out because of the exclusion criteria were recruited.

6.3.5 Inclusion criteria for controls

Controls were selected from the same baseline population that yielded the cases. Up to five controls were selected per case, as this increased the power of the study.[100] Cases and controls were matched on sex, age and GP practice. Inclusion criteria for controls consisted of:

- Being registered with the same GP practice as their matched case when the latter was diagnosed with cancer
- Being aged ≥40 years when their case was diagnosed with cancer
- Having at least one entry in the clinical or referral files (see below, Section 6.5.2) in the year prior to cancer diagnosis in the case

- Being alive on the date on which their case was diagnosed with cancer

6.3.6 Exclusion criteria for controls

Exclusion criteria for controls consisted of:

- Exclusion of their matched case for any reason
- Being diagnosed with the same cancer as their matched case before 1 January, 2000
- Being diagnosed with the same cancer as their matched case after 1 January, 2000 – as this would mean they were eligible to be a case themselves
- Being a control for another case

6.3.6.1 Subanalysis of bladder cancer dataset

In separate sub-analyses, the bladder cancer dataset was restricted to just those cases with diagnostic codes included in ICD category C67. In other words, cases (and their matched controls) were dropped if they had received a diagnosis of carcinoma *in situ* or of neoplasm of uncertain behaviour of bladder. This was carried out in order to estimate the risk of bladder cancer in symptomatic patients commensurate with the post-1998 ICD definition of the disease. The risk estimate results of this analysis are reported in Appendix 6: Risk estimates for the post-1998 definition of bladder cancer (C67).

6.4 Matching

Cases were matched on sex, age and GP practice, using observation window matching, with up to five controls. Conventionally, matching is used to control for confounding. In CAPER studies, confounders would have: (1) a causative association with the disease (albeit 'caused by' rather than 'causing') and (2) a non-causative association with the risk marker under investigation. It is hard to envisage a framework for confounding between two risk markers, given their dependence on causative pathology. Equally, it is reasonable to assume that there is no mechanism by which a risk factor could have a non-causative association with a risk marker, since risk factors exert their effect *before* cancer develops whereas risk markers only become apparent *after* cancer is established. Therefore, smoking, alcohol consumption and obesity – all commonly considered confounders in conventional case–control studies – do not fulfil these criteria in CAPER studies. The purpose of matching here was not to control for confounding; rather, it was to minimise bias as described below.[100]

6.4.1 Matching on sex

Matching on sex was carried out to mitigate any reporting bias caused by differential consulting patterns and behaviour between men and women. A study of consultation rates in England between 1995 and 2008 using QResearch® data reported that women tended to visit the GP more frequently than men between the ages of 19 and 60 years. Aggregated data on crude consultation rates indicated that men aged 15–19 years attended twice per

year, rising to 5 times at the age of 55–59 years, whereas attendances by women at the same time points were just under 5, rising to nearly 7. In contrast, there was no observable difference in attendance patterns between men and women in the very young or very elderly; for example, in 2008 the highest median practice consultation rates were 12.9 (IQR: 10.2–16.1) per person-year for men aged 85–89 years and 12.5 (IQR: 9.9–15.2) per person-year for women of the same age.[101] A study of nearly 4 million patients in 2010 in The Health Improvement Network (THIN) database reported that overall consultation rates in primary care were 30% lower for men than women. The size of the difference varied with age, being greatest among 21- to 39-year-olds (ratio of male to female consultations = 0.40, 95% CI: 0.39–0.40). Of more relevance to CAPER studies, in patients aged >40 to 57 years, the male : female ratio was 0.62 (95% CI: 0.62–0.63), and in those older than 58 any difference had largely disappeared (0.92, 95% CI: 0.91–0.93).[102] The same authors then investigated consulting patterns in the 24 months prior to a diagnosis of colorectal or lung cancer or malignant melanoma. Participants included all patients over 16 years permanently registered in THIN with an incident diagnosis of lung (*n* = 11,081) or colorectal (*n* = 12,189) cancer or malignant melanoma (*n* = 4,352) between 1997 and 2006. Interestingly, the authors found no evidence that consulting patterns varied between men and women, although it should be noted that they recorded just the number of consultations, investigating neither their content nor their nature.[103] Nevertheless, matching on gender remains prudent given that the propensity for women to consult more than men is retained in the control group.

6.4.2 Matching on age

As reported above, consultation rates vary with age, reflecting how disease – and therefore symptom – profiles change throughout life.[101] The latest statistics summarised by Cancer Research UK indicate that roughly half of all incident cases of both pancreatic and bladder cancer were in people older than 75,[53,61] whereas the median age of over 40s that consult their GP was reported to be 62 years.[104] Therefore, without matching on age, cases would be an average of >10 years older than controls, leading to comparisons being made between two groups with different symptom profiles.

To protect against patient identification, the CPRD did not release exact dates of birth, only the year. For the purposes of matching, all participants were assumed to have been born on 1 July as this falls midway through the year.

6.4.3 Matching on GP practice

Matching on GP practice was used as a proxy to control for the known dependence of consulting patterns on socioeconomic status and location. A study of 60 GP practices in England and Wales including half a million patients reported that a higher proportion of people in social classes IV and V consulted their GP compared with those in classes I and II. Consultation rates also varied with geographical location.[105]

Matching on GP practice was also done to avoid potential recording bias due to variations in the documentation of clinical information between GP practices. All

data submitted to the CPRD are subjected to stringent quality controls and the majority of participating GP practices use ViSion software to generate the electronic health record (see Section 2). In addition, GP practices are strongly encouraged to follow ViSion recording guidelines.[3] All of this, in theory, minimises between-practice variation in data collection. Nevertheless, some variation in adherence to the recording guidelines will remain, not least because ViSion can be customised; for example, by creating thesauri of Read codes that appear automatically when GPs start to type in certain symptoms. In addition, GPs themselves will vary, not only in terms of what they decide to include in the notes but also how they record it – as a Read code or in the text. While the CPRD variable 'staffid' allows identification of the practice staff member who entered the data, it does not necessarily follow that this person is the actual GP. Therefore, as a pragmatic compromise, matching on GP practice was carried out to minimise recording bias between participants in the study.[100]

6.5 Data collection and data work-up

6.5.1 Data collection

Data were collected by GPs as part of routine primary care. The mechanism of data recording is described in detail in Section 2.2 and will not be repeated here.

6.5.1.1 Reporting bias

Matching on sex, age and GP practice all minimise the potential for reporting bias. One could also argue, however, that patients in the early presentation stage of cancer might be starting to become worried about their health and have a lower threshold than controls for attending the GP. In the original CAPER studies, this was examined by looking for differences in attendance for fractures as a proxy for general attendance patterns between comparison groups. This was chosen as fractures were thought to be equally likely across the entire study population, given that incident rather than metastatic cancer was investigated. Limitations of this approach are discussed in the next section (Section 6.5.1.2).

Reporting bias may also occur owing to differences between what the patient tells the GP and what the GP hears and records. This is discussed below in Section 6.5.1.2, Recording bias.

6.5.1.2 Recording bias

The two main potential sources of recording bias in observational studies of data recorded as part of everyday clinical care are discussed below.

6.5.1.2.1 Consistency of recording between comparison groups

Unbiased estimates of clinical features in cases and controls require that GPs are equally likely to record information reported by the two comparison groups. Indeed, the assumption made by all researchers using CPRD data is that GPs

have no systematic tendency to alter the information that they record according to a patient's ultimate diagnosis. However, when making a diagnosis, GPs will test alternative hypotheses iteratively when taking the history. They will focus on those symptoms reported by the patient that fit with their working diagnosis, and use these to inform the tests and referrals they decide to request. This heuristic approach – in effect 'framing' by the working diagnosis – will influence which information reported by the patient is subsequently recorded by the GP.[106] [107,108,109,110] This calls into question the validity of the above assumption about consistency of recording style between comparison groups, because the cases are shortly to be diagnosed with cancer whereas the controls are free of the disease. It is conceivable that preference is given to recording those signs and symptoms that justify the clinical management, introducing recording bias in accordance with the clinical significance of a symptom for the cancer under investigation.

6.5.1.2.2 Use of codes versus text

In my study, unbiased estimates of clinical features in cases and controls also require that the accessibility of information within the electronic medical record is the same for cases as for controls. The assumption made by all researchers using CPRD data is that GPs have no systematic tendency to alter their recording style according to a patient's ultimate diagnosis. There is evidence that paper-based systems are more amenable than computerised systems to the recording and quantification of symptoms.[11] By analogy, it is conceivable that GPs may find it easier to record symptoms in the 'comments box' rather

than as a Read code. Differential use of text-only recording between cases and controls would introduce a recording bias because, as described above (see Section 2.5), Read codes are readily and routinely available to researchers using CPRD data, whereas the text is generally withheld. Therefore, symptom information recorded only in text and not as a code will be 'hidden' to conventional studies and result in false-negative assignment of patients to symptom-free status.

This was tested in the original CAPER studies by studying fracture rate and so far no evidence of bias has been unearthed. However, the choice of fractures may be criticised because they are unambiguous diagnoses requiring treatment and follow-up in fracture clinic. Arguably, a fracture is unrepresentative of symptoms or signs suggestive of more than one possible diagnosis that may or may not require referral. It is possible that recording styles may differ in these two scenarios. Therefore, this study specifically addresses the validity of assumptions made about GPs' recording style.

6.5.2 Raw data

Raw data were received from the CPRD in the generic American Standard Code for Information Interchange (ASCII) format, on multiple CD-ROMs. In total, 12 file types were provided, of which 7 (6 of coded and 1 of text data) were used. Their contents are described briefly below.

6.5.2.1 Matching files

The CPRD provided a complete list of potential cases and controls in separate 'matching' files for each cancer site. Each file consisted of three variables, the first two being unique identifiers for cases (capatid) and for controls (copatid), respectively; the third, an ordinal number that took a value of 0 for cases and of between 1 and 5 for controls.

6.5.2.2 Patient files

The patient files contained basic demographic data and patient registration details. Data were held in 21 variables, including a unique patient identifier (patid, derived from capatid and copatid in the matching files), gender and year of birth and an indicator stating whether the patient's data met quality standards.

6.5.2.3 Practice files

The practice files contained details about the GP surgeries that contribute to the CPRD, including their region and collection information. Data were held in four variables, including a unique identifier for each practice (pracid) and the up-to-standard date, i.e. the date at which the practice data were last deemed to be of research quality.

6.5.2.4 Clinical files

These files contained the entire medical history, including clinical signs, symptoms, diagnoses and deaths, entered into the GP system using Read

codes. Data were stored in 11 variables, including: the unique patient identifier patid; the event date; the medcode identifying the clinical event[a]; and a text identifier (textid) to allow linkage of the medcode with its paired free text record. Patients may have had more than one row of data.

6.5.2.5 Referral files

These files contained referral details recorded on the GP system. Information pertained to patient referrals to external care centres (normally to secondary care, such as hospitals for inpatient or outpatient care), and included speciality and referral type. Data were held in 14 variables including patid, event date, textid, medcode and others indicating the referral speciality.

Clinical and referral files were regarded as essentially equivalent, so they were appended into a master 'clinical and referral' dataset of medcodes.

6.5.2.6 Test files

These files contained the patients' test results recorded in the GP system. The data were held in 13 core variables, plus a variable number of data fields for the test results themselves. The data were coded using a Read code, chosen by the GP, identifying the type of test used. Data were stored differently according to the test that was carried out and denoted either qualitative text-based (for

[a] Mapped 1:1 by the CPRD from the Read code, purely to facilitate data handling.

example 'Normal' or Abnormal') or quantitative (i.e. a numeric value that needs to be compared to the normal range – also provided) results.

6.5.2.7 Text files

The text files were not routinely available to CAPER studies, but were specifically requested for this investigation. Text fields in the medical records of participants in the bladder[4] and pancreatic[5] cancer studies were searched for evidence that the patient attended the GP for abdominal pain, jaundice or haematuria. The CPRD were asked to extract all the following word strings:

- abdominal pain: 'bdominal pai' or 'bdo pai' or 'ain in abdo' or finally 'pigastric pai' – if searches were not case sensitive, the obvious first letter was added
- jaundice: 'aundice' or 'cterus' or 'cteric'
- haematuria: 'aematuria' or 'lood in urine'

CPRD staff examined the strings and redacted any information that could identify the patient. String variables were created, consisting of the extracted search term plus the three words either side to aid interpretation.[a]

[a] The CPRD advised that limiting the accompanying text to three words either side of the search term would be the most cost-effective way of obtaining interpretable data (the data are charged by the word).

Data were held in four variables; namely, patid, pracid, textid and the text itself. Patients may have had more than one row of data.

Visual inspection suggested that sorting of extracts into the appropriate symptom categories at the CPRD had introduced some errors; for example, some extracts relating to haematuria were sorted to the abdominal pain group. Therefore, all extracts were combined into a single dataset and re-sorted into the appropriate category by identifying the original search terms given to the CPRD. Some extracts contained search terms for more than one symptom, and so duplicate copies were sorted into each symptom dataset.

6.5.3 Variables extracted from the raw data

The variables extracted from the raw data files and used in this study are listed in Table 6.1.

6.5.4 Raw data processing and work-up

The raw data files supplied by the CPRD were imported individually into Stata (version 11 in the original studies; version 13 in this PhD). The variables listed in Table 6.1 were extracted and used to construct the final dataset.

This section describes the raw data processing and work-up carried out in readiness for dataset construction in CAPER studies.[4,5] Not all steps were repeated when re-creating the datasets for this study; instead, some variables from the original studies were imported. Such variables are clearly identified

and explanations given as to why it was necessary to do this rather than re-create them from scratch.

Table 6.1 Variables extracted from the raw data files

Raw data file	Variables used	
Patient file	Unique patient identifier	Year of birth†
	Gender	Date of death†
Clinical file	Unique patient identifier	Event date
	Medcode	Additional clinical details code
	Unique text identifier	
Referral file	Unique patient identifier	Unique text identifier
	Medcode	Event date†
Test file	Unique patient identifier	Test value
	Medcode	Lower limit of normal range
	Event date†	Upper limit of normal range
Text fileᵖ	Unique patient identifier	The text string#
	Unique text identifier	
Matching file	Unique case identifier‡	Ordinal within patient group‡
	Unique control identifier‡	

Notes:
ᵖData clean-up required.
†Raw data provided in string format and required conversion to elapsed date format for analysis in Stata.
#Raw data provided in string format and required conversion to binary variable for analysis in Stata.
‡Raw data used to derive variables essential for analysis of matched data.

Most of the variables in the data files could be used as provided, while others required conversion to a format suitable for analysis in Stata, and this is also described below.

6.5.4.1 Text file data clean-up

The validity of each observation within its symptom group was checked by searching for and identifying the original search terms sent to the CPRD. Incongruous observations were reassigned to their correct symptom dataset, after which duplicate observations were identified and surplus records dropped.

The text records were supplied as sequences of characters in ‘string’ variable format and were converted to binary variables (symptom present or absent) suitable for analysis. This was a lengthy process and unique to my PhD; therefore, the details are described in a separate section (see Section 6.6).

6.5.4.2 Deriving variables essential for analysis of matched data

The study population contained a maximum of 5,000 patients with a particular cancer, plus up to five matched controls per case. Each case and their controls formed a 'group'. The data in the matching files (see Section 6.5.2.1) were used to create variables that uniquely identified each patient (patid), their case/control status and the group to which they were assigned. These variables were essential for analysis of matched data. The patient identifier variable, patid, was similarly generated by the CPRD and is the same as that referred to throughout Section 6.5.2.

6.5.4.3 Patient age

A variable ‘age’ was generated, assuming that all participants were born on 1 July, as this falls midway through the year.

6.5.4.4 Date of diagnosis

The diagnosis date was taken to be the first occurrence in the patient’s electronic medical record (the ‘clinical and referral’ master) of a medcode listed in the cancer’s diagnostic thesaurus (see Appendix 2: Disease thesauri). Diagnosis dates were imported from the original studies, because these determined the period over which text records were examined for evidence that the patient was attending with haematuria, jaundice or abdominal pain.

6.5.4.5 Age at diagnosis

Age at diagnosis was calculated, as were binary variables that assigned patients to the following age groups at diagnosis: >60 and ≤60 years.

6.5.4.6 Determining the analysis period within patient groups

The aim of analysis was to identify how cancer presents in the year before diagnosis. Therefore, all controls were assigned their case’s diagnosis date, in order to anchor the end point of the analysis period within patient groups. To ensure that data were restricted to the analysis period, all events in the ‘clinical and referral’ master dataset occurring either after, or more than a year before, a patient group’s diagnosis date were dropped. Duplicate observations were identified and surplus copies deleted. The resulting dataset was saved and

used as the source of *coded* evidence of possible features of cancer, as described in Section 6.8. Data in the test files were similarly restricted to the analysis period within patient groups.

The CPRD created a 'redundant' date of 1/1/2500 and any observations assigned this date were dropped from the study. Also dropped were any observations whose date of recording was not provided.

6.6 Construction and application of the algorithm to convert text strings to binary variables

6.6.1 Introduction

Pre-analysis work-up was required to convert the strings in the text files (see Sections 6.5.2.7 and 6.8.2) to a format suitable for analysis. Binary variables were generated indicating whether or not each text record included evidence that the patient had attended for abdominal pain, haematuria (visible or non-visible) or jaundice in the year prior to diagnosis in the case. Reference to a symptom in an observation should not be interpreted as confirmation that the patient has the clinical condition in question; indeed, doctors will also document when particular signs and symptoms are absent to demonstrate that they have taken a thorough history.[111]

6.6.2 Using an algorithm to interpret the meaning of text strings

CPRD databases can include thousands of observations. Reading and interpreting each one by hand is time consuming to the point of impracticality;

furthermore, it renders the process vulnerable to inconsistencies in decision-making. To enhance the accessibility and utility of the text data, a 'triage' algorithm was devised to increase the automation of this process in a way that was reliable and also quantified the uncertainty inherent in the resulting variables.

6.6.2.1 Theory of negation

Arguments in the algorithm were underpinned by the theory of negation in natural language, which is a complex area and touched on only very lightly here. For an in-depth discussion of the structure, use and meaning of negation in natural language, I referred to *A Natural History of Negation.*[112] The type of negation applicable here is 'contradictory opposition', which has two distinguishing criteria[112]: it applies to pairs of statements and the pairs are mutually exclusive, i.e. one must be true and the other false. This fits well with the nature of symptoms, which are also binary entities in that they are either present or they are not at the time of the consultation.

6.6.2.2 Construction of the algorithm

The arguments of the algorithm were derived from the rules governing sentence construction (syntax) as applied to the theory of negation. Full details are given in Appendix 3: Algorithm construction. Code for the algorithm that converted the text strings to discrete variables was created in Stata (version 13).

A preliminary assessment of the data revealed that binary classification, 'symptom present' and 'symptom absent', was not sufficient, and that a third

category ‘meaning unclear’ was required, at least in the interim. Sometimes this was because of uncertainty over timing and sometimes because the text was nonsensical (some examples are shown in Table 6.2).

Table 6.2 Examples of text strings relating to haematuria

String of text	Interpretation
AFTERNOON HAS HAD HAEMATURIA AND NOT FEELING	Symptom positive
CLOTS BLOOD IN URINE	Symptom positive
PAINFUL HAEMATURIA; MSU PENDING; TEL	Symptom positive
. HELPFUL IN STOPPING HAEMATURIA IN THE PAST	Unclear – uncertainty over timing of haematuria
TAKEN. NO SOME HAEMATURIA OVER LAST 1/52	Uninterpretable
HAS HAD NO HAEMATURIA. SHE HAS HAD	Symptom negative

Therefore, the decision was taken to classify each observation initially into one of the following groups:

1. ‘**Symptom negative**’ – used to describe those observations in which symptoms were explicitly or implicitly described as absent at the time the patient consulted the GP.

2. ‘**Symptom positive**’ – used to describe those observations in which symptoms were contemporary.

3. '**Meaning unclear**' – used to describe when the patient's status regarding the symptom could not be ascertained with certainty.

Identification of those observations whose meaning was unclear was difficult to automate since there are few rules on which to base decision-making. I decided to search for trigger words and highlight those observations as potentially unclear (see Appendix 3: Algorithm construction).

6.6.3 Application of the algorithm

Semi-automated classification of observations was carried out to reduce the number requiring manual assessment:

1. All observations were initially classed by the algorithm as symptom-positive.

2. The algorithm identified the negation of symptoms and reclassified those observations as 'symptom negative'.

3. The algorithm identified potentially ambiguous observations, based on the presence of trigger words, and reclassified them as 'meaning unclear'. These observations required later manual assessment to see if they could be reclassified as either symptom-negative or symptom-positive.

4. Observations that were not picked up by the algorithm in steps 2 and 3 remained classified as symptom-positive by default. This was later verified by manual assessment.

The accuracy of this process was assessed by comparing the output with that of a specially constructed reference standard (see Section 6.7).

6.6.4 Generation of variables for analysis

The classified text strings were converted to binary variables suitable for analysis. Unclear observations were deemed to provide insufficient evidence that the symptom was present, and so were reclassified as ‘symptom absent’ to achieve binary classification.

Variables were created for abdominal pain, jaundice and the two types of haematuria – visible and non-visible.

6.6.5 Missing data

Missing data occurred where the text fields did not contain any mention of the symptom in question. The lack of either reporting or recording of the symptom was interpreted as no evidence that the symptom occurred.

Missing data may also arise from misspelling and typographical errors, as affected text strings would not have been picked up by the CPRD’s search criteria. To get a measure of the potential for this type of error, I exported a random sample of the text record received from the CPRD into Word and proofread them. The rate of spelling and typographical errors was reported.

6.7 Validating the classification procedure using the diagnostic test model

Text data are inherently noisy and it was important to assess the accuracy with which observations were categorised, to get a measure of classification error. In the diagnostic test model, accuracy is reported in terms of sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) in comparison with a gold standard. In this context, the gold standard would be the true meaning of the free text being categorised.

6.7.1 Applicability criteria of the diagnostic test model

For the diagnostic test model to be applicable to the semi-automated procedure for classifying free text observations, several criteria need to be fulfilled[113]:

1. The final classification system must be binary.

2. A gold standard must be used to assess the free text observations and to establish the true category of each one. (The classification procedure's output is then compared with that of the gold standard in a contingency table. As the gold standard provides error-free classification, all false-positives and false-negatives can be attributed with certainty to error in the classification procedure.)

3. The same gold standard is used to assess all observations.

4. Assessments made using the gold standard and the semi-automated classification procedure should be made at the same time, to eliminate changes that occur over time.

Criteria 3 and 4 were readily met and did not pose a problem in this scenario; however, the same cannot be said of criteria 1 and 2.

6.7.1.1 Criterion 1: alternatives when classification is not binary

As described above, we had the problem of observations whose meaning remained unclear. There were three possible approaches to dealing with this, as described below.

6.7.1.1.1 Receiver operating characteristic curve

Where ordinal categories have been derived from underlying numerical variables, observations can be classified into a binary system based on a threshold or cut-off value. The performance of the threshold value can be quantified by plotting sensitivity against (1 – specificity), to give a receiver operating characteristic curve.[114] This method can be applied to string data by assigning a numerical score to each observation, such as is used in information retrieval systems. For example, using a Bayesian inference network approach, Aranow[115] developed an *ad hoc* method for classifying radiology reports according to the level of evidence that mammograms belonged to one of three categories regarding the presence of suspicious conditions: positive, negative or uncertain. It was proposed that if the system accurately categorised

mammograms as positive or negative, only those categorised as uncertain would require manual review.

Developing a similar system for scoring our free text observations and setting cut-off values to enable a binary classification system would not be trivial and so alternatives were explored.

6.7.1.1.2 Correction for misclassification of binary exposure variables

Correction for misclassification of binary exposure variables is complex because the errors, unlike those on numerical variables, correlate with the true value of the binary variable. In essence, the error can only take a value of 0, −1 or +1: it is always 0 when the classification is correct, but when the classification is incorrect the error is +1 when the true value is 0, and −1 when the true value is 1.[114] Making repeated measurements is suggested as a way of imputing missing values, but this option was not available in our study.

6.7.1.1.3 Sensitivity analysis

In the absence of an observation scoring system, a binary classification system can be imposed on data. This is achieved by taking the view that there is insufficient evidence to classify ‘unclear’ observations as ‘symptom positive’, instead assigning them to the category ‘symptom negative’. Subsequent sensitivity analysis would be required, first merging the ‘uncertain’ category with the ‘positive’ category, and then with the ‘negative’ category, to create the following binary classification systems:

a. Symptom negative + Symptom not negative (i.e. merged groups ‘symptom positive’ and ‘meaning unclear’).

b. Symptom positive + Symptom not positive (i.e. merged groups ‘symptom negative’ and ‘meaning unclear’).

Calculating the metrics sensitivity, specificity, PPV and NPV for each of these two groups would give their maximum and minimum values, with the true value for our data set lying somewhere in between.

Sensitivity analysis was the preferred option in this study.

6.7.1.2 Criterion 2: alternatives when there is no gold standard

The absolute gold standard for interpreting the free text would be to ask the GP who wrote the note to review the entire free text content and confirm the meaning. Clearly, this was impossible, so we had to seek an alternative.

6.7.1.2.1 Manual assessment by a single expert

Some authors have taken manual assessment by an expert as a proxy for the gold standard of free text interpretation.[116,117,118] For example, Wang used manual assessment by a single medical practitioner blinded to the output of an algorithm written to classify documents in a fully automated process.[118] However, this does not allow any assessment of reliability (i.e. the reproducibility of the measurement), which is particularly important in this setting as the interpretation of free text observations is so subjective. Therefore, alternatives were explored.

6.7.1.2.2 Constructing a reference standard

The evaluation of index tests when there is no gold standard was the subject of a review by Rutjes.[113] Of the four main methods identified, constructing a reference standard using the panel of consensus method was chosen as the most appropriate. The methods used to achieve this are now described.

6.7.2 Constructing and validating the reference standard

A panel of two raters was formed, both practising GPs and both supervisors of my thesis. A random sample of 100 observations about haematuria was extracted from the text files using Stata.

6.7.2.1 The pilot trial

The text strings were exported to Excel and sent to each rater independently. Using a drop-down list, each rater was asked to select the category they felt most accurately reflected the meaning of the text: symptom present, symptom absent or meaning unclear. The raters worked independently, and no conferring was allowed.

I collated the results, and identified and reported the level of agreement between the two raters. I chose to use the weighted kappa statistic, as it gives a measure of agreement over and above that due to chance alone. The formula used to calculate the quadratic weights (w) was:

$w_{ij} = 1 - [(i - j)^2/(k - 1)^2]$,

where *i* and *j* are the individual categories and *k* is the total number of categories (i.e. 3).

'Symptom present' was denoted as category 1; 'meaning unclear', category 2; and 'symptom absent', category 3. The quadratic weights were calculated as:

$w_{11} = w_{22} = w_{33} = 1 - [(0)^2/ (3 - 1)^2] = 1$ (i.e. complete agreement)

$w_{12} = w_{21} = 1 - [(1)^2/ (3 - 1)^2] = 0.75$ (i.e. partial agreement)

$w_{13} = w_{31} = 1 - [2^2/ (3 - 1)^2] = 0$ (i.e. complete disagreement)

The kappa score is known to be sensitive to marginal inhomogeneity, being reduced by symmetrical imbalance and enhanced by asymmetrical imbalance.[119] Therefore, the marginal totals were calculated and reported as an indication of the degree of homogeneity.

6.7.2.2 Category definitions

The results of the pilot trial were shared with both raters and definitions agreed for each category (see Section 6.6.2.2).

6.7.2.3 Finalising the reference standard

The raters were then asked to re-assess the same random sample of 100 observations, independently, according to the agreed definitions. The raters subsequently discussed the discrepancies, in a bid to maximise agreement in the reference standard to near gold standard status. If observations remained

stubbornly indecipherable by both raters, or the raters disagreed as to their meaning, their classification as unclear was retained. The percentage defined in this category was reported.

6.7.2.4 Validating the reference standard

The reference standard was validated to provide a measure of its potential as a source of error, by reporting final levels of inter-rater agreement using the kappa statistic.[113] While the interpretation of the kappa statistic is variable (see Table 6.3), there is general agreement that a value >0.8 indicates very good or almost perfect agreement. A kappa value of 1 would indicate that a true gold standard had been established.

Table 6.3 Interpretation of the kappa statistic

Kappa value	Interpretation of strength of agreement by Altman[120]	Kappa value	Interpretation of strength of agreement by Landis[121]	Kappa value	Interpretation of strength of agreement by McGinn[122]
<0.20	Poor	<0.40	Moderate or poor	0	None
				0–0.2	Slight
0.21–0.40	Fair			0.2–0.4	Fair
0.41–0.6	Moderate	0.4–0.75	Fair to good	0.4–0.6	Moderate
0.61–0.80	Good	>0.75	Excellent	0.6–0.8	Substantial
0.81–1.00	Very good			0.8–1.0	Almost perfect

6.7.2.5 Finalising the reference standard

To finalise the reference standard, observations were dropped if agreement could not be reached as to their meaning, or the raters agreed that they were uninterpretable. The weighted kappa was calculated and reported.

6.7.3 Validating the classification process

The purpose of constructing the reference standard was to validate the classification process, to provide a measure of its potential as a source of misclassification of free text observations.[113] To that end, the classification process was used to categorise the same random sample of observations as was used to construct the reference standard. Validation was reported in terms of the weighted kappa statistic, as above, along with marginal totals.

Summary

Having cleaned-up the raw data and converted the text strings to variables suitable for analysis, using a validated classification method, I moved on to identifying evidence that patients attended their GP in the analysis period for risk markers of cancer.

6.8 Identifying records of attendance for possible features of cancer

This section describes how both coded and text records of attendance for possible features of cancer were extracted from the patients' electronic records.

Recap

Figure 2.1 is repeated here as a timely reminder of what I mean by ‘coded’ and ‘text’ records. During the consultation, GPs are obliged to select a Read code first to make a clinical entry in the patient’s record (the coded record), after which they are free to write in the comments box (the text record). The text may or may not relate to the Read code.

Figure 2.1 (repeated) Screenshot showing the comments box that becomes available (red arrow) once a Read code has been selected in ViSion

The CPRD subsequently map the alphanumeric Read codes 1:1 to numeric medcodes, purely to facilitate analysis. As discussed earlier (see Section 2.5), the majority of studies restrict their analysis to the coded record. However, in this study, the text record was also examined for evidence that patients attended for abdominal pain, haematuria or jaundice (see Sections 5 and 6.5.2.7).

6.8.1 Identifying coded records of patient attendance for symptoms possibly indicative of cancer

A three-step process was used to extract coded records of patient attendance for all those symptoms identified in the original studies as being the most strongly associated with cancer.[4,5] This involved creating a separate thesaurus of medcodes for each possible symptom of cancer, and identifying when medcodes from a thesaurus had been recorded in the electronic medical record (the 'clinical and referral' master, see Section 6.5.2.5).

Step 1 – finding synonyms for potential features of cancer

Step 1 was included to maximise data capture and involved finding synonyms for each potential feature of cancer. A variety of resources were used, the main one being the latest version of the coding system International Classification of Primary Care, ICPC-2.[123] This classification system was chosen because, compared with Read codes, it has greater emphasis on symptoms. Other resources consulted included literature from patient support groups, the internet (e.g. medical education websites, patient information websites and discussion forums) as well as amenable medical practitioners.

Step 2 – creating symptom thesauri for each potential feature of cancer

In step 2, the descriptions of all medcodes were searched for any occurrence of synonyms identified in step 1. A shortlist of medcodes containing these synonyms was saved as a Stata dataset. The final list of pertinent medcodes

was selected manually from the shortlist to create a symptom thesaurus. Codes were omitted if they pertained only to treatment, to diagnostic procedures or to a family or past history. Codes for diseases characterised by the symptom of interest were considered carefully; for example, a thesaurus for 'diarrhoea' might include 'gastroenteritis'. Each symptom thesaurus was agreed with the primary investigator and practising clinician, Professor Willie Hamilton.

When re-creating the datasets for my PhD, the bespoke symptom thesauri of the bladder[4] and pancreatic[5] cancer studies were re-used to ensure the accurate identification of information that had been missed because it had been recorded solely in text.

In a separate strand of analysis, I wanted to compare how symptoms were recorded across the bladder and pancreatic cancer datasets. For this I could re-use the haematuria and jaundice symptom thesauri from the bladder and pancreatic cancer studies, respectively, as these symptoms were not common to both cancers. However, abdominal pain was a shared feature, for which each study had derived a unique symptom thesaurus.[a] While both included medcodes for general abdominal pain, the bladder cancer thesaurus included codes for features such as abdominal tenderness, rigidity and colic. In contrast, the pancreatic cancer thesaurus had included codes for epigastric pain, indigestion and dyspepsia. Therefore, I used the generic symptom thesaurus for

[a] This 'customising' was done to ensure that each thesaurus was optimised for the synonyms described in step 1, to match the pattern of abdominal pain experienced in each cancer.

abdominal pain created by the research group for the distinct purpose of comparing how abdominal pain was recorded across both cancer sets.

When creating the generic symptom thesaurus, the definition of abdominal pain was deliberately kept tight. Tenderness was not considered as a synonym for pain, because it is elicited on examination and is classified as a sign rather than a symptom.[91] Discomfort was also omitted as a synonym for pain, because it was thought to be at or below the lower limit of the spectrum of pain severity, which, as discussed above, is very subjective and unreliable (see Section 4.5.3.2). I could have used a Delphi process to select the codes for inclusion in the thesaurus; however, I did not have unlimited access to my clinical supervisors' time, and chose to prioritise their input for the classification of text records over the choice of symptom codes. Furthermore, a Delphi process was not used in the selection of codes for the jaundice and haematuria thesauri, so its introduction would have been inconsistent with the methodology of the original studies.[4,5]

All symptom thesaurus codes are reported in Appendix 4: Symptom thesauri.

Step 3 – identifying if and when the patients' records contained symptom thesaurus codes

Step 3 identified if and when symptom thesaurus codes were recorded in the patients' records. This was done by merging the symptom thesaurus with the master dataset of clinical and referral files (see Sections 6.5.2.4 and 6.5.2.5)

and extracting those observations that were common to both to a new dataset of *coded* records of patient attendance for that symptom.

The individual codes used by the GP to record the symptom were noted, after which a single binary variable was created simply to denote that there was a coded record of patient attendance for the symptom in the analysis period.

The dataset was truncated just to three variables, one each for patient identification, patient attendance for the symptom and the event date. Data were excluded if the date of recording was not available. Multiple records of a symptom on the same day were assumed to refer to a single event; therefore, surplus entries were dropped to avoid duplicate reporting. In the dataset, patients had more than one row of data – one for every date on which they presented with the symptom.

This process was repeated for each symptom separately and individual datasets were later combined to construct the final dataset.

6.8.2 Identifying text records of attendances for haematuria, abdominal pain or jaundice

A three-step process was used to identify where haematuria (visible or non-visible), abdominal pain or jaundice was recorded solely in the text and not as a code.

Step 1: Identifying the event date and the medcode paired with the text record

In ViSion, the option to enter information as text is only available once a Read code has been selected; therefore, every text record will have a matched Read code that is subsequently mapped to a medcode.

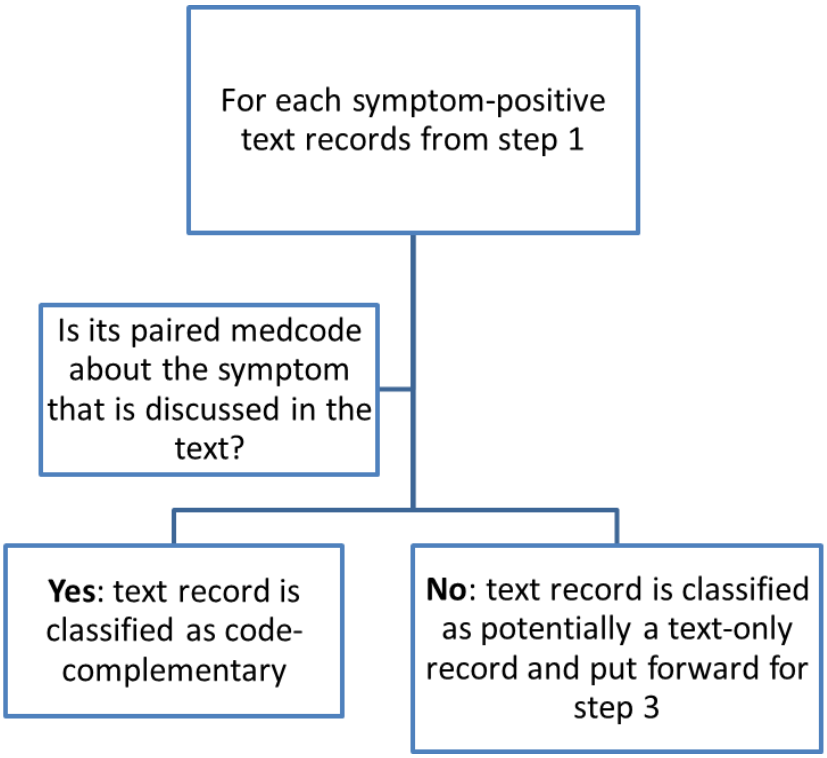
All text records that had been classified as symptom-positive (see Section 6.6.3) were put forward for Step 1.

Step 1 identified the Read Code and medcode paired with each text observation and the date of recording (see Figure 2.1, repeated above, for a reminder of how GPs enter clinical information). The text records in the raw text files for each of haematuria, abdominal pain and jaundice (see Section 6.5.2.7) could be uniquely identified by a combination of two variables: the patient identifier (patid) and the text identifier (textid). This allowed accurate mapping of each text record to its paired medcode and event date in the electronic medical record (the 'clinical and referral' master dataset of medcodes, see Sections 6.5.2.4 and 6.5.2.5).

Step 2: Identifying whether the paired medcode is related to the symptom discussed in the text

Step 2 looked at the relationship between the text record and its paired medcode. If the paired medcode was related to the symptom discussed in the text (see Figure 6.1), the text record was classed as code-complementary; if not, it was considered to be a potentially text-only record.

Figure 6.1 Step 2: Discriminating between code-complementary and potentially text-only records

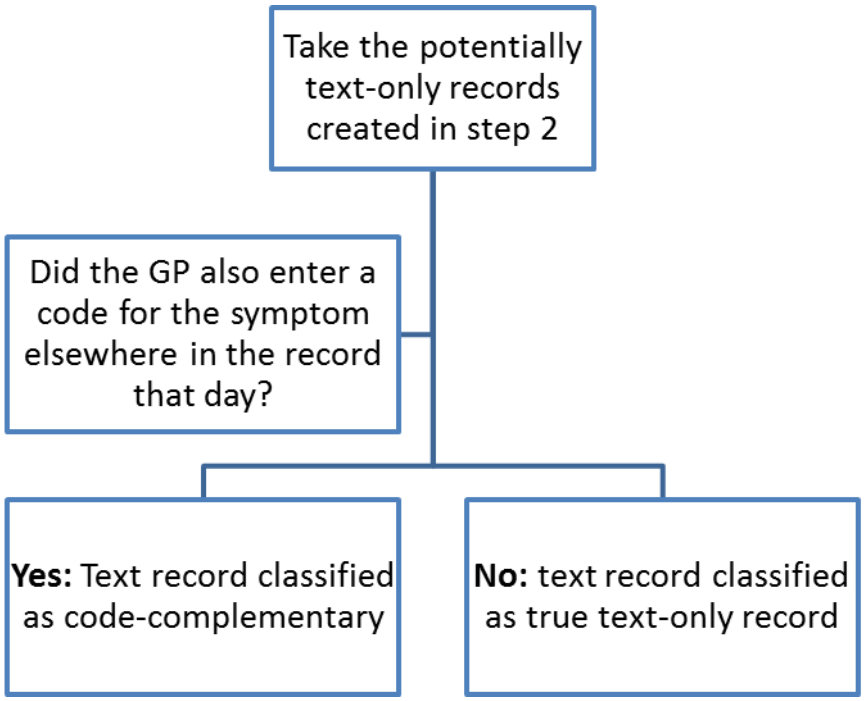


To do this, the dataset of symptom-positive text records with their paired medcode created in step 1 was merged with that symptom's thesaurus of medcodes. Text records that matched with a medcode in the symptom thesaurus were classified as code-complementary. Observations that could *not* be matched to a medcode in the symptom thesaurus were classified as potentially text-only and put forward for step 3.

Step 3: Identifying true text-only recording

This step checked to see if each potentially text-only record produced in step 2 was the sole record of the patient's attendance with the symptom, or whether the GP had in fact also used a code elsewhere in the notes (see Figure 6.2).

Figure 6.2 Step 3: Identifying true text-only recording



To do this, for each symptom, the dataset of potentially text-only records created in step 2 was combined with that of the coded records of attendance (see Section 6.8.1) into a single file. Observations from the two datasets were matched on date. Potentially text-only records that were made on the same day as a coded record of attendance were classified as code-complementary, whereas unmatched observations were classified as true text-only records. Only true text-only records were retained for analysis.

To illustrate steps 1–3, Table 6.4 lists some examples of text records, their paired medcode, date of recording as well as the intermediate and final classification.

Table 6.4 Examples of free text records for a single patient and their classification

	Example number:			
	1	**2**	**3**	**4**
Event date	01/06/2010	31/06/2010	31/06/2010	07/07/2010
Free text comment	Presenting complaint was abdominal pain [a]	Ongoing problems with abdo pain [b]	Ongoing burning upper abdominal pain [c]	Summary: Abdo pain [d]
Paired medcode	177	68	177	56
Description	Abdominal pain	Chest infection	Abdominal pain	Seen in GP surgery
Classification in Step 1	Code-complementary	Potentially text only	Code-complementary	Potentially true text only
Final classification in Step 3	Code-complementary	Code-complementary	Code-complementary	True text only

[a] Free text record in column 1 was classified as code-complementary, because the paired medcode was also for abdominal pain.

[b] Free text record in column 2 was classified as potentially text-only in step 2, because the paired medcode was *not* for abdominal pain, but a search of the whole record identified a separate medcode on the same day (31/06/2010) for abdominal pain (column 3), so the final classification in step 3 was code-complementary.

[c] Free text record in column 3 was classified as code-complementary, because the paired medcode was also for abdominal pain.

[d] A true text-only record, because a search of the whole record could not identify any codes for abdominal pain on that day for that patient.

6.8.3 Identifying records of abnormal investigation results

In the original studies, a three-step process was used to find coded records of abnormal investigation results for study participants in the analysis period. In

brief, this involved creating thesauri of medcodes for investigations related to cancer, looking up the results in the test files, and generating a binary variable to indicate whether the result was abnormal or not.

I replicated this process when re-creating the datasets for my PhD, as it was a useful learning exercise and increased my understanding of the data.

Step 1: creating 'test' thesauri

Step 1 consisted of creating 'test thesauri' similar to symptom thesauri discussed above (see Section 6.8). A list of all blood tests relevant to the cancer under investigation was drawn up by the primary investigator and practising GP, Professor Willie Hamilton. Then all medcodes relating to these tests were identified and saved as separate 'test thesaurus' datasets, one for each test type. In the pancreatic cancer study, all hepatic enzyme tests were merged into a single composite variable. Similarly, tests for erythrocyte sedimentation rate, plasma viscosity and C-reactive protein were collated into a single variable called 'raised inflammatory markers'.

Step 2: getting the test results

Step 2 consisted of extracting the test results from the test files (see Section 6.5.2.6) and quantifying the amount of information recorded by each medcode in the thesaurus. The 'core' medcodes were noted, i.e. those that collectively contributed 90% of the results. Any test medcode that contributed <1% of the total test results was dropped from the thesaurus as it was deemed unrepresentative. Next, the test value was compared with the normal range,

which was provided in the test file. If the normal range for core medcodes was missing, the value was imputed using the modal value for the test. If the normal range for a non-core medcode was not provided, the data were not used.

Step 3: identifying abnormal test results

Step 3 involved creating a binary variable denoting whether the patient had an abnormal test result.

6.9 Variables

This section describes the variables that were created for the features of cancer.

6.9.1 Variables created from the coded records

Categorical and binary variables were created to replicate those of the original studies, for all features of cancer, using the coded records; for example, `haematuria`, `abdominal pain` and `jaundice`.

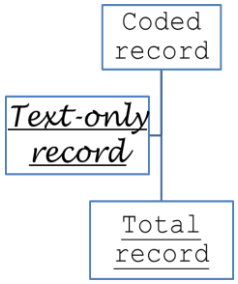
6.9.1.1 Binary variables

Binary variables were created to identify patients who had attended at least once for haematuria, jaundice or abdominal pain during the analysis period. These variables were used to estimate the positive likelihood ratio (see Section 6.12.3.4) and the positive predictive value (see Section 6.12.3.5) according to the coded record.

Note about variables

Several variables for features of cancer were created to allow full investigation of the research questions outlined in Section 5. I have used fonts to clarify the recording method used (see Figure 6.3).

Figure 6.3 Coded records were supplemented with text-only records to create variables reflecting the entire content of the electronic medical record



`Courier New` denotes variables generated from codes, i.e. `haematuria`, `abdominal pain` and `jaundice` – these replicate the variables from the original study (and may or may not have a code-complementary text record).

Underlined *Lucida Handwriting* is used for variables generated from the text only, i.e. *haematuria*, *abdominal pain* and *jaundice*. This is used to represent events that were *never* recorded in code.

Underlined `Courier New` is used for the total record, where the variable is based on the code but is supplemented with records made solely in the text, e.g. `haematuria` is the sum of `haematuria` plus *haematuria*.

6.9.1.2 Binary variables

Binary variables were created to identify patients who had attended at least once for haematuria, jaundice or abdominal pain during the analysis period. These variables were used to estimate the positive likelihood ratio (see Section 6.12.3.4) and the positive predictive value (see Section 6.12.3.5) according to the coded record.

6.9.1.3 Discrete, categorical variables

Categorical variables were created that identified each patient's attendance patterns for features of cancer where the event was recorded using a code. They took the form: 0: no attendance; 1: only one attendance; 2: only two attendances; …; *n*: *n* attendances or more, where *n* was determined as the last category that contained at least 5% of cases. In practice, the categorical variables were often equivalent to the binary variables described above, because it was rare for more than 5% of cases to attend for a feature more than once in the analysis period.

The categorical variables were used as explanatory variables in the conditional logistic regression to get a measure of the association between attendances for these risk markers and a cancer diagnosis, as estimated from coded records. Where numbers permitted, these variables were also used to investigate whether re-attendance for a feature was clinically significant in terms of having different odds of a cancer diagnosis.

6.9.2 Variables created from the text-only records

For the symptoms haematuria, abdominal pain and jaundice, categorical and binary variables were created from the text-only records.

6.9.2.1 Binary variables

The binary *haematuria*, *abdominal pain* and *jaundice* variables identified those patients who attended at least once for the symptom in the analysis period, but whose *first five attendances* for the symptom were made using the text only, i.e. the GP never used a code to record any of these attendances.

This variable was used to identify under-reporting in the original studies of the *total number* of patients who attended for the symptom at least once in the year prior to the cancer diagnosis, which also relates to research questions 1 and 5 (see Section 5).

6.9.2.2 Discrete, categorical variables

The categorical *haematuria*, *abdominal pain* and *jaundice* variables identified the attendance patterns for haematuria, abdominal pain or jaundice for those patients whose *first five attendances* were recorded in text only.

The categorical variables were used as explanatory variables in the conditional logistic regression to get a measure of the association between attendance for the risk marker and cancer, as estimated from text-only records. This relates to research questions 1 and 5 (see Section 5).

Where numbers permitted, these variables were also used to investigate the clinical significance of re-attendance for a feature.

6.9.3 From the combined text-only and coded records

Binary and categorical variables – haematuria, abdominal pain and jaundice – were created using coded records supplemented with individual attendances recorded solely in the text (see Figure 6.3, above).

6.9.3.1 Binary variables

The binary variables identified the total number of patients who attended at least once for the symptom in the analysis period, as identified from coded or solely text-only records.

These variables were used to re-estimate the positive likelihood ratio and positive predictive values for cancer in patients presenting with haematuria, abdominal pain or jaundice in light of all the information recorded by the GP about that patient. This relates to research questions 2 and 5 (see Section 5).

6.9.3.2 Discrete, categorical variables

The categorical variables identified the complete attendance pattern for a symptom whether it was recorded using a code or in the text. These variables were used as explanatory variables in conditional logistic regression. The association between attendance for a symptom and cancer was re-estimated in light of all the information recorded by the GP about that patient. This relates to

research questions 2 and 5 (Do the text data provide additional value to coded data?).

6.9.4 Recording style variable

For those patients who did attend at least once in the analysis period with haematuria, abdominal pain or jaundice, a binary variable was created to indicate the recording style used overall. ‘Coded’ meant that the GP had used a code for at least one of the attendances, whereas ‘text-only’ meant that the GP had always used text and never a code. This variable was used to examine whether there is an association between recording style and the type of symptom, the patient status or the context of presentation. This analysis relates to the following research questions:

1. Are studies of coded data vulnerable to bias arising from the differential use of text and codes between comparison groups?

2. Does recording style vary with type of symptom?

3. Does the recording style vary with the context of presentation of a symptom?

6.10 The final dataset

6.10.1 Creating a baseline patient demographic dataset

A baseline patient demographic dataset was created, in which there was one line of data per patient. This included the variables age, case/control status,

diagnosis date, gender and age at diagnosis. Finally, the binary variable that identified excluded patients was imported from the original studies, to ensure that I was studying exactly the same participants.

6.10.2 Addition of variables for signs, symptoms and investigation results

To construct the final dataset, the baseline demographic dataset was appended with each of the datasets containing evidence that patients attended in the analysis period for symptoms, and evidence of their investigation results. Before they were appended, the datasets were converted so that there was a single row of data per patient.

Duplicate recordings made on the same day were identified and surplus observations were dropped.

6.11 Missing data

The General Medical Council requires that GPs document all important findings, both positive and negative, to create a record that ensures continuity of care.[97] Therefore, the assumption was made that a clinical event only occurred if there was evidence that it did. Consequently, missing data were interpreted as the equivalent of the absence of a sign or symptom, or that a test result would have been normal had it been carried out.

6.12 Analysis

6.12.1 Threshold for inclusion of clinical features

In the original studies, only symptoms, signs and test results that occurred in at least 5% of the cases or controls were studied. The frequency was examined in cases and in controls because, arguably, features presented preferentially by controls may have been protective for cancer. However, given that the study aims to identify risk markers, this seems relatively unlikely. In practice the features were all identified in the cases.

To some extent this cut-off of 5% was arbitrary and requires justifying. At one extreme, a very low threshold increases the chances of identifying associations that, while statistically significant, are not clinically meaningful. This is akin to making a type I error, detecting an effect that is not actually present. However, raising the threshold too far the other way would increase the risk of missing those features that were clinically important, if not terribly frequent; in other words, making a type II error of failing to detect an effect that is present. A cut-off of 5% was thought to represent a pragmatic balance between these two extremes.

6.12.1.1 Using text data to examine the impact of a 5% threshold

Rather fortuitously, inclusion of text data allowed the impact of applying this 5% threshold to be examined. Text records about haematuria were subdivided into those referring to its visible and non-visible forms (see Section 4.5.1.2). In the

original bladder cancer study, non-visible haematuria was recorded too infrequently (in 127/4,915 cases = 2.6%)[a] to warrant inclusion in analysis.[4] Therefore, we took the opportunity to assess whether supplementing with text-only records would increase the number of cases presenting with non-visible haematuria over the 5% threshold. If this turned out to be true, this would enable the first assessment of the risk of bladder cancer in patients presenting with non-visible haematuria. This was an exciting diversion, though still germane to the overall study question on the value of free text data. It was particularly timely, as the National Institute for Health and Care Excellence (NICE) were eager for evidence on this specific issue as they prepared to revise their guidelines on referral for cancer.[57]

6.12.2 Outcome measures in event-level analysis

6.12.2.1 Recording style preferences

The preference for symptom recording style at the event level – coded (± a complementary text record) or text only – was reported regardless of case or control status. This analysis was designed to address research question 3: does recording style vary with type of symptom?

[a] The investigations file was also searched for laboratory results of urine dipstick tests for blood (medcode 14561) and for red blood cells visible on urine microscopy (medcode 38204). No such tests had been carried out, confirming that this is an investigation carried out within GP surgeries, and not in the laboratory.

The total number of attendances was reported within cancers for alarm symptoms (visible haematuria for bladder cancer; jaundice for pancreatic cancer) and for non-alarm symptoms (abdominal pain for both cancers, non-visible haematuria for bladder cancer) separately.

6.12.2.2 Estimating diagnostic intervals

These analyses relate to research question 5: Do the text data provide additional value to coded data?

The diagnostic interval is estimated as the number of days between first attendance for a symptom and the date of cancer diagnosis (DI_{coded}). Studies that restrict their analysis to codes may underestimate the diagnostic interval, because it is possible that the GP recorded the symptom solely in the text at an earlier consultation.

Analysis was restricted to cases (true bladder and pancreatic cancer patients) and diagnostic intervals were estimated for abdominal pain (bladder and pancreatic cancer datasets), haematuria (bladder cancer) and jaundice (pancreatic cancer).

Determination of whether the diagnostic intervals followed a normal or asymmetrical distribution was made in several ways: first, using a box and whisker plot of the data; secondly, by regressing the quantiles of the data against the quantiles of a normal distribution with the same mean and variance as the sample being tested; and, thirdly, using the D'Agostino K-squared test of normality.[124]

Three approaches were taken to assess whether text data provide additional value to coded data when it comes to estimating diagnostic interval.

First, the earliest coded record of attendance was used to estimate DI_{coded} for haematuria and abdominal pain in the bladder cancer dataset, and for jaundice and abdominal pain in the pancreatic cancer dataset. The coded records of attendance were augmented with text-only records and the diagnostic interval was re-estimated using the earliest recorded attendance for the symptom, be it recorded using a code or solely in the text ($DI_{coded/text}$). Some patients had both a coded and a text record, some had only a coded record and some only a text record. Therefore, there is no statistical test suitable for testing whether the diagnostic interval was altered by the addition of text-only records, because the data were neither fully matched nor completely unmatched. Therefore, just the summary statistics are reported as the median interval (in days) plus the 25% to 75% interquartile range.

Secondly, in matched analysis, the sign rank test for matched data was used to test the null hypothesis, H_0, that $DI_{coded} - DI_{text/coded} = 0$. To fulfil the requirement for matched data, analysis was restricted further, to those cases who had both coded and text records of attendance for the symptom. For example, they may have attended three times during the analysis period, and their first attendance was recorded in code, the second using both methods and the third attendance solely in the text. For each patient the diagnostic interval was estimated from the earliest coded record (DI_{coded}) and from the earliest record, which may have been coded or in the text ($DI_{text/coded}$). The summary statistics are reported as

the median interval (in days) plus the 25% to 75% interquartile range. The significance level was set at $p<0.05$.

Thirdly, in unmatched analysis, the Wilcoxon rank-sum (also known as Mann-Whitney) test was used to test whether estimates of diagnostic intervals from coded (DI_{coded}) and from text (DI_{text}) records are drawn from the same population. To fulfil the requirement for unmatched data, analysis was restricted to those with solely text records and to those with solely coded records. The summary statistics are reported as the median interval (in days) plus the 25% to 75% interquartile range. The significance level was set at $p<0.05$.

6.12.3 Outcome measures in patient-level analysis

6.12.3.1 Recording style preferences

The preference for recording style at the patient level – coded (± a complementary text record) or text only[a] – was reported regardless of case or control status for alarm and non-alarm symptoms separately to address research question 3: does recording style vary with type of symptom?

The total number of attendees was reported within cancers for alarm symptoms (visible haematuria for bladder cancer; jaundice for pancreatic cancer) and for

[a] Note: Each patient was categorised by the style used to record their attendances for the symptom. 'Coded' was assigned if *any* record of a symptom was in coded form; conversely, 'text-only' was designated only when *all* instances were noted solely in the text (see Section 6.9.4).

non-alarm symptoms (abdominal pain for both cancers, non-visible haematuria for bladder cancer).

The numbers of patients with a history of attendance for haematuria, jaundice or abdominal pain who were overlooked in the original studies because their records were lost in the hidden text were reported.

6.12.3.2 Comparison of symptom recording style between cases and controls

Evidence of an association between patient status (case or control) and recording style (coded or text-only record) was obtained using the $\chi 2$ test to address research question 2: are studies of coded data vulnerable to bias arising from the differential use of text and codes between comparison groups? For each symptom examined, the total number of patients attending at least once in the analysis period was determined from codes and text-only records. The proportion of the total attributable to each recording style was determined in cases and controls separately and displayed in a contingency table as well as graphically. The $\chi 2$ test was carried out for each symptom within each cancer dataset to test the null hypothesis that there is no association between patient status and recording style. The significance level was set at $p<0.05$.

6.12.3.3 Test of the association between context of symptom presentation and recording style

The association between context of symptom presentation and recording style was tested. Methods described in Sections 6.12.3.1 and 6.12.3.2 were repeated

for visible haematuria (the alarm symptom for bladder cancer) in the pancreatic dataset, and for jaundice (the alarm symptom for pancreatic cancer) in the bladder cancer dataset.

As described in Section 4.5.3, the characteristics of abdominal pain vary with the site of origin. For this reason, the original studies each created a bespoke abdominal pain symptom thesaurus. Therefore, a generic symptom thesaurus for abdominal pain was created to allow comparison of recording of this 'low-risk but not no-risk' symptom across both datasets. These tests were carried out to address research question 4: does the recording style vary with the clinical context of presentation of a symptom?

6.12.3.4 Likelihood ratios

The likelihood ratio is a way of estimating the performance of a diagnostic test. In CAPER Studies, the positive likelihood ratio summarises how many times more likely cases are to have a clinical feature compared with controls. In other words, it is the probability that cases have a feature divided by the probability that controls have the same feature. It is calculated mathematically as $\frac{p(S+|D+)}{p(S+|D-)}$, where $p(S+|D+)$ is the probability of symptom presence ($S+$) given (|) that cancer is present ($D+$) and $p(S+|D-)$ is the probability of symptom presence given that cancer is not present. This calculation equates to sensitivity / (1− specificity).

A likelihood ratio greater than 10 is considered as providing strong evidence of an association.[125]

Confidence intervals for the positive likelihood ratio were calculated in Excel using the following equation:

$$\exp(\ln\frac{p_1}{p_2} \pm \sqrt{\frac{1-p_1}{p_1 n_1} + \frac{1-p_2}{p_2 n_2}})$$

Where, p_1 is the sensitivity (number of cases with the feature, divided by the total number of cases)

p_2 is 1−specificity (where specificity is the number of controls without the feature divided by the total number of controls)

n_1 is the total number of cases

n_2 is the total number of controls

6.12.3.5 Positive predictive value

The positive predictive value (PPV) is another measure of diagnostic test performance. It estimates the probability that someone has a disease given that they have a positive test result, *p* (disease | positive test result), and is expressed as a percentage 'chance'. The PPV was first applied to cancer diagnostics in primary care by Holtedahl in 1990,[126] and was developed extensively by Hamilton.[4,5,40,41,42,43,44,45,46,127,128] NICE recently lowered the

threshold for the PPV of symptoms that should trigger a cancer assessment from approximately 5% to 3%.[9,129]

Arithmetically, the PPV can be calculated as the number of 'true positives' divided by the 'total number of positives' identified by the test, i.e. the sum of true positives and false positives. Unlike specificity and sensitivity, which remain fixed, arithmetically derived PPV varies with disease prevalence. This is because, unless the test's specificity is 100%, decreasing prevalence must be accompanied by an increasing number of false-positives since the total number of test results remains constant. Consequently, it is inappropriate to estimate PPV arithmetically in case–control studies, which by their very nature arbitrarily fix the prevalence as well as the balance between cases and controls. The adjustment of prevalence occurs in two ways. First, the population is enriched in that there are generally many more cases in the study than would be expected in the population from which the cases derive. Secondly, there may be many controls per case.

Instead, PPV was estimated using Bayes' Theorem (see Appendix 5: Derivation of positive predictive value using Bayes' theorem for full derivation and example for illustration):

Posterior odds = (prior odds × *f*) × positive likelihood ratio (1)

The adjustment factor, *f*, was necessary because the likelihood ratio and the prior odds were each representative of different populations; namely, the consulting and the whole populations, respectively. Therefore, the prior odds

were adjusted to reflect the value in the consulting population to produce risk estimates that are meaningful to GPs. The adjustment factor, *f*, derived from the proportions of cases and of controls who consulted their GP during the period of the study:

$$f = \frac{\frac{x_c}{x_t}}{\frac{y_c}{y_t}} \qquad (2)$$

Where x_c = the number of eligible cases who consulted in the study sample, x_t = the total number of cases in the study sample, y_c = the number of controls who consulted and y_t = the number of controls overall.

The assumption made here was that any patient who has not consulted the GP in the year prior to diagnosis did not have a positive indication of disease, i.e. their test results would have been normal or they did not attend for the symptom whose PPV was being estimated.

Finally, the posterior odds were converted back to a probability to estimate the PPV:

PPV = posterior odds / (1 + posterior odds) (3)

6.12.3.6 Odds ratios

In the case–control studies, I estimated the strength of association between symptoms and a cancer diagnosis according to the ratio of the odds of a symptom in cases compared with the odds of the same symptom in controls.

For both univariable and multivariable analysis, odds ratios were estimated by Stata using conditional logistic regression to accommodate the matched study design and binary outcome. Using categorical explanatory variables, I estimated the odds ratio for specific numbers of attendances for each symptom in the analysis period; for example, only once, only twice and three or more.[130]

I include below a short explanation of how odds ratios for cancer in patients with given symptoms are calculated manually. This will be referred to later on in the Discussion of how the odds ratio estimate is affected by bias that derives from the omission of text records from analysis.

In case–control studies, to estimate the association between a symptom and cancer it is first necessary to calculate the probability of that symptom in the cases and controls separately, i.e. on the conditional presence of cancer. The conditional probabilities can be written using the same notation as that used in the derivation of PPV using Bayes' Theorem (see Appendix 5: Derivation of positive predictive value using Bayes' theorem), where D is the disease (cancer) under investigation and S is the symptom. Furthermore, $D+$ indicates that cancer is present (i.e. cases) and $D-$ that it is not (i.e. controls); similarly, $S+$ and $S-$ indicate symptom presence and absence, respectively. The lack of either reporting or recording of the symptom was assumed to mean there was no evidence that the symptom occurred (see Section 13.2.6 for Discussion of missing data). The conditional probabilities are easily displayed in a contingency table; for example, the probability of symptom presence given

(denoted by '|') that the patient has cancer (i.e. is a case) is represented as *p*(S+|D+):

		Outcome	
		Symptom present	**Symptom absent**
Exposure	**Cancer (cases)**	*p*(*S*+|*D*+)	*p*(*S*−|*D*+)
	No cancer (controls)	*p*(*S*+|*D*−)	*p*(*S*−|*D*−)

Knowing that odds = probability / (1 – probability) and that symptoms can only be present or absent, we can write:

Odds of symptom presentation in cases = $\frac{p(S+\mid D+)}{p(S-\mid D+)}$

Odds of symptom presentation in controls = $\frac{p(S+\mid D-)}{p(S-\mid D-)}$

Therefore, the ratio of odds in the cases compared with the controls is:

Odds ratio = $\frac{p(S+\mid D+)\times p(S-\mid D-)}{p(S-\mid D+)\times p(S+\mid D-)}$, which is simply the **cross-product of the contingency table**.

Therefore, the odds ratio is vulnerable to bias in the estimates of symptom recording between cases and controls.

6.12.4 Regression analysis

6.12.4.1 Univariable analysis to establish independent association between risk markers and cancer

In the original studies, univariable analysis was used as the first step to modelling the presentation of cancer in primary care. It was carried out for each potential feature to assess whether it had an independent association with cancer. The z test was used to test whether the odds ratio was equal to 1 (i.e. no association) and the threshold p value for retention was $p<0.1$ to ensure that important variables were not omitted.

It was not necessary to repeat univariable analysis in this study for all potential features of cancer. Rather, it was reserved only for those features included in the final models published in the original studies and for the new text-based variables.

6.12.4.1.1 Post-estimation tests

In my univariable analysis of the new text-based variables, the Wald test was used for each symptom to test the strength of its association with a cancer diagnosis. The level of significance was set at $p<0.1$ as in the original studies.

To address research question 5 ‘Do the text data provide additional value to coded data?’ the Wald test was used to test whether code- and text-based variables gave similar estimates of the association between cancer and any of

the features abdominal pain, haematuria or jaundice.[131] The significance level was set at $p<0.05$.

Subsequently, code- and text-based variables were combined into a single, composite variable for regression analysis (see Figure 6.3 above).

Where 5% of cases attended at least twice in the analysis period, the Wald test was used to test whether re-attendance was clinically significant in terms of association with cancer. The significance level was set at $p<0.05$.[131]

6.12.4.2 Multivariable analysis to model the main features of cancer presentation

The models derived in the original studies were adopted by this study; therefore, the processes carried out to determine the main features of cancer presentation were not replicated here. In brief, in the original studies, variables that passed univariable analysis were grouped on clinical grounds. Within each group, stepwise regression was carried out to select those features with the strongest association with cancer. The threshold value for retention was $p<0.05$ (z test). Surviving variables were submitted to a final round of stepwise regression. The most significant features were retained in the final model using $p<0.01$ (z test).

6.12.4.3 Replication of the final model

The final models of the original studies were replicated and post-estimation Wald tests were used to test whether re-attendance with a feature altered the odds ratio for cancer. The significance level was set at $p<0.05$.[131]

6.12.4.4 Comparison of models with and without text-based variables

The final models of the original studies were re-created using coded variables `haematuria`, `jaundice` and `abdominal pain`. The final models were compared with new models supplemented with text-only variables for *abdominal pain* (both cancers), and *jaundice* (pancreatic) or *haematuria* (bladder) (likelihood ratio test). This corrected for the underestimation of the number of patients presenting at least once with the symptom in the analysis period. If inclusion of text-based variables was significant, the coded variables `haematuria`, `jaundice` and `abdominal pain` were replaced with the composite variables `haematuria`, `jaundice` and `abdominal pain`.

This analysis was carried out to investigate research question 5: do the text data provide additional value to coded data?

6.12.4.5 Post-estimation tests

Either the Wald or the likelihood ratio test could have been used as a post-estimation test of whether the addition of text-based explanatory variables improved the fit of the model. There appears to be no clear indication as to which is superior, although many statisticians seem to prefer the likelihood ratio

test when both are suitable. Therefore, the likelihood ratio test was used to test nested models (one with and one without the text-based variables) in this study.[131] The significance level was set at $p<0.05$.

6.12.4.6 Testing for effect modification

Evidence for effect modification was sought on clinical grounds; for example, between symptoms and diseases that were associated both with one another and the cancer – such as urinary tract infection and haematuria. This allowed us to ascertain the strengths of association between a risk marker and cancer separately for each level of the effect modifier. Evidence of effect modification is useful clinically, as it suggests to GPs what other information they should elicit when taking a history in order to assess the risk of cancer properly.

Careful consideration was given to whether it was useful clinically to look for effect modification between the visible and non-visible forms of haematuria in the bladder cancer study. As discussed above (see Section 4.5.1.2), they are both defined as an abnormal presence of blood in the urine and lie on a continuum of one disease. Indeed, patients cannot have both visible and non-visible forms of the disease simultaneously. Therefore, it was decided to be inappropriate to look for effect modification between these two forms of haematuria.

6.12.4.7 Post-estimation tests

A post-estimation Wald test was used to test for the significance of effect modification and the significance level was set at $p<0.05$.

6.12.5 *Post-hoc* analysis: modelling the outcome 'text-only recording' of visible haematuria in the bladder cancer dataset

6.12.5.1 Introduction

Finally, as *post-hoc* analysis, the outcome 'text-only recording' was modelled for visible haematuria using likely underlying causes (benign and malignant) as explanatory variables.

When a patient consults in primary care because of visible haematuria, the GP has to consider a wide range of possible causes – from the benign through to the serious, including malignancy (see Section 4.5.1.5). Having considered the patient's history and other presenting features, the GP must then decide how to record the consultation in the patient's medical record.

My study provided an opportunity for *post-hoc* analysis to test the hypothesis that GPs tend to use codes to record clinically significant events (e.g. indicative of cancer) and text for anything perceived not to be worrisome (e.g. benign and self-limiting pathology). This hypothesis arose following discussion of my results for visible haematuria with two of my supervisors (Professor Willie Hamilton and Dr Kevin Barraclough) and visiting academic, Dr Peter Hjertholm, all of whom are practising GPs – hence the *post-hoc* nature of the analysis.

I tested for associations between the outcome 'text-only recording of visible haematuria at the patient level' and possible causes of visible haematuria, i.e.

benign (UTI, stones) and serious (bladder cancer). The modification of this association by gender was also explored, because UTIs are more common in women than men, while bladder cancer is more common in men than women.

This *post-hoc* analysis also fits in well with research question 4: Does the recording style vary with the clinical context of presentation of a symptom?

6.12.5.2 Mixed-effects logistic regression

In this part of the analysis, the outcome variable was not case/control status; rather, it was whether a patient's entire attendance for visible haematuria was made solely in the text, using the binary variable *visible haematuria*. This outcome was modelled using a mixed-effects logistic regression model adjusting for the random effect of clustering due to non-independence of observations within GP practices (as a result of matching). Choice of potential explanatory variables was based on possible causes of the symptom, which are discussed in Section 4.5.1.5. Visible haematuria in the cases was assumed to be attributable to bladder cancer, whereas visible haematuria in the controls was assumed to have a benign underlying pathology. Therefore, control–case status was used as a proxy for malignant vs benign cause. Other benign causes investigated included urinary tract infection and calculi.

Evidence for possible reasons underlying each symptom was sought using the methods described in Section 6.8, including the compilation of medcode thesauri that were agreed with the lead investigator.

Univariable analysis was carried out to determine whether there was an independent association between each possible cause of the symptom and the outcome of text-only recording. Successful variables were submitted to multivariable analysis and stepwise regression was used to select the final model.

6.12.5.3 Effect modification

Effect modification by gender of the association between patient status (as a proxy for benign vs malignant cause of haematuria) and text-only recording of haematuria was investigated, because bladder cancer is more common in men than women.

7 Results: study participants

7.1 Bladder study participants

The participants of the bladder cancer study and their characteristics were determined in the original study.[4] While the results in Subsection 7.1 are not strictly part of my PhD, they are presented at this point to enable complete interpretation of my extension of the original study.

7.1.1 Bladder cancer study cases

In total, 4,935 potential cases were identified as having a clinical or referral record of incident bladder cancer, as defined by the list of 20 Read codes agreed between the primary investigator (Professor Hamilton) and the CPRD research team (see Appendix 2: Disease thesauri).[a]

7.1.1.1 Excluded cases

After application of the exclusion criteria, 20 potential cases (*n* = 19 men, *n* = 1 woman) were excluded from the study, for reasons given in Table 7.1. There were no cases who had not consulted their GP in the year prior to their diagnosis, leaving a total of 4,915 cases in the study.

[a] Using the pre-1998 definition of bladder cancer.

7.1.2 Bladder cancer study controls

In total, 24,098 patients were identified by the CPRD as eligible for consideration as controls.

7.1.2.1 Excluded controls

After application of the exclusion criteria, 2,380 potential controls (*n* = 1,958 men, *n* = 422 women) were excluded from the study for reasons given in Table 7.1.

Table 7.1 Bladder cancer study exclusions

Case or control	Reason for exclusion	Number of patients excluded
Case	No matched control identified	13
	Metastatic cancer present	7
	Subtotal	*20*
Control	Diagnosed with bladder cancer after the year 2000	125
	Diagnosed with bladder cancer before the year 2000	134
	No data recorded in their medical record in the analysis period	2,086
	Matched to a case who was excluded because they had metastatic cancer	35
	Subtotal	2,380
Total		**2,400**

7.1.3 Bladder cancer study matching

After exclusions, there were 4,915 bladder cancer cases matched to 21,718 controls on age, sex and GP practice, as reported in Table 7.2. The majority of cases were matched to at least four controls.

Table 7.2 Bladder cancer study matching

No. (%) of cases with:					
1 control	**2 controls**	**3 controls**	**4 controls**	**5 controls**	**Total**
57 (1.2)	160 (3.3)	419 (8.5)	1,311 (26.7)	2,968 (60.4)	4,915

7.1.4 Characteristics of the bladder cancer study participants

7.1.4.1 Bladder cancer study participant demographics

The age and gender of study participants are reported in Table 7.3.

Table 7.3 Bladder cancer study participant demographics

Age group[a]	Case (*n* = 4,915)		Control (*n* = 21,718)	
	Male, n (%)	**Female, n (%)**	**Male, n (%)**	**Female, n (%)**
<60 years	417 (8.5)	140 (2.8)	1,631 (7.5)	639 (2.9)
≥60 years	3,146 (64.0)	1,212 (24.7)	13,821 (63.6)	5,627 (25.9)
Total	*3,563 (72.5)*	*1,352 (27.5)*	*15,452 (71.1)*	*6,266 (28.9)*

[a]The age group refers to the age at diagnosis in the case.

Overall, 3,563 of the 4,915 (72.5%) cases were men and 1,352 (27.5%) were women, reflecting the higher incidence of bladder cancer in the male population. Matching on gender ensured similar proportions in the control group: 15,452 (71.1%) men and 6,266 (28.9%) women. The median age at diagnosis in the cases and in the matched controls was 74 (interquartile range: 66–80) years.

7.2 Pancreatic cancer study participants

The participants of the pancreatic cancer study and their characteristics were determined in the original study.[5] While the results in Subsection 7.2 are not strictly part of my PhD, they are presented at this point to enable complete interpretation of my extension of the original study.

7.2.1 Pancreatic cancer study cases

In total, 3,647 potential cases were identified as having a clinical or referral record of incident pancreatic cancer, as defined by the list of 25 Read codes agreed between the primary investigator (Professor Hamilton) and the CPRD (see Appendix 2: Disease thesauri).

7.2.1.1 Excluded cases

After application of the exclusion criteria, 12 potential cases were excluded from the study, for reasons given in Table 7.4. There were no cases who had not consulted their GP in the year prior to their diagnosis, leaving a total of 3,635 cases in the study.

7.2.2 Pancreatic cancer study controls

In total, 17,977 patients were identified by the CPRD as eligible for consideration as controls.

7.2.2.1 Excluded controls

After application of the exclusion criteria, 1,518 potential controls were excluded from the study for reasons given in Table 7.4.

Table 7.4 Pancreatic cancer study exclusions

Case or control	Reason for exclusion	Number of patients excluded
Case	No matched control identified	2
	Tumour not originating in the pancreas	10
	Subtotal	*12*
Control	Diagnosis of pancreatic cancer	64
	No data in the year before diagnosis	1,414
	Case excluded	40
	Subtotal	*1,518*
Total		**1,530**

7.2.3 Pancreatic cancer study matching

After exclusions, there were 3,635 cases matched to 16,459 controls on age, sex and GP practice, as reported in Table 7.5.

Table 7.5 Pancreatic cancer study matching

No. of cases with:					
1 control	**2 controls**	**3 controls**	**4 controls**	**5 controls**	**Total**
26 (0.7)	90 (2.5)	251 (6.9)	840 (23.1)	2,428 (66.8)	3,635

7.2.4 Characteristics of the pancreatic cancer study participants

7.2.4.1 Pancreatic cancer study participant demographics

The age and gender of pancreatic cancer study participants are reported in Table 7.6. Overall, 1,743 of the 3,635 (48.0%) cases were men and 1,892 (52.0%) were women. Matching on gender ensured similar proportions in the control group: 7,627 (46.3%) men and 8,832 (53.7%) women. The median age at diagnosis in the cases and in the matched controls was 73 (interquartile range: 65–80) years.

Table 7.6 Pancreatic cancer study participant demographics

Age group[a]	Case (*n* = 3,635)		Control (*n* = 16,459)	
	Male, n (%)	**Female, n (%)**	**Male, n (%)**	**Female, n (%)**
<60 years	320 (18.4)	241 (12.7)	1,227 (16.1)	1,105 (12.5)
≥60 years	1,423 (81.6)	1,651 (87.3)	6,400 (83.9)	7,727 (87.5)
Total	*1,743 (48.0)*	*1,892 (52.0)*	*7,627 (46.3)*	*8,832 (53.7)*

[a]The age group refers to the age at diagnosis in the case.

8 Results: text string classification

The accuracy and validity of the methods used to convert text strings to binary variables were assessed by comparing the final classification of a random sample of observations with that of the reference standard. Section 8.1 describes the results relating to construction of the reference standard. The performance characteristics of the classification algorithm itself are reported later in Section 8.2. Other sources of uncertainty in text variables are reported in Section 8.3. Finally, the results of text string classification are reported in Section 8.4.

8.1 The reference standard

As described in the Methods (see Section 6.7), the reference standard was created in four stages: an initial one in which two raters independently classified a random sample of 100 text strings (the pilot study); a second stage where the raters clarified the criteria underpinning their decisions; and a third step in which the same sample was re-assessed. In the final stage, text strings were dropped if the raters could not agree to their meaning, or agreed that they were wholly uninterpretable.

8.1.1 The pilot study results

The results of independent assessments of the random sample of 100 text strings carried out by the two raters in the pilot study are displayed in Table 8.1.

The marginal totals for the three categories were symmetrically unbalanced: for both observers the prevalence of 'current' haematuria' (74% for rater 1 and 77% for rater 2) greatly exceeded that of either 'no current haematuria' (12% for rater 1 and 15% for rater 2) or 'meaning unclear' (14% for rater 1 and 8% for rater 2).

Table 8.1 Results of independent assessments of the random sample of 100 text strings

		Rater 1			Total
		Current haematuria	**Meaning unclear**	**No current haematuria**	
Rater 2	**Current haematuria**	**68**	7	2	77
	Meaning unclear	4	**2**	2	8
	No current haematuria	2	5	**8**	15
Total		74	14	12	100

The observed proportion of agreement between the raters was (68 + 2 + 8)/100 (78%); however, this measure overestimated the true level of agreement as some of this will have arisen purely by chance. The chance-corrected weighted Kappa was 0.7 (95% CI: 0.6–0.9). In terms of inter-rater agreement, this is interpreted as fair to good by Landis and Koch,[121] as good by Altman[120] and as substantial by McGinn *et al.*[122] The symmetrical imbalance within the sample in favour of 'current haematuria' means that the reported chance-corrected weighted Kappa is likely to underestimate the level of agreement.

Overall, raters 1 and 2 agreed that 2% of the observations were wholly uninterpretable.

8.1.2 Clarification of category definitions

As described in Section 6.6.2.2, most disagreements related to uncertainty over whether the record was referring to a symptom that was historical or contemporary; although, in some instances, the text was nonsensical (some examples were shown in Table 6.2).

For clarity, the finally agreed definitions are repeated here:

'**Symptom negative**' – used to describe those observations in which symptoms were explicitly or implicitly described as absent at the time the patient consulted the GP.

'**Symptom positive**' – used to describe those observations in which symptoms were contemporary.

'**Meaning unclear**' – used to describe when the patient's status regarding the symptom could not be ascertained with certainty.

8.1.3 Results after agreement of category definitions

After discussion and agreement of the definitions for the three categories, the two raters independently re-assessed the same random set of observations. The results are reported in Table 8.2. The anticipation was that fewer

observations would be classed as ‘meaning unclear’. Rater 1 did tend to assign this category less frequently, with its percentage reducing from 14% to 3%. However, the use of this category by rater 2 remained fairly steady at 8% and then 11%.

Table 8.2 Results of independent re-assessment of the random sample of 100 text strings

		Rater 1 (revised)			Total
		Current haematuria	**Meaning unclear**	**No current haematuria**	
Rater 2 (revised)	**Current haematuria**	**78**	1	4	83
	Meaning unclear	2	**2**	7	11
	No current haematuria	0	0	**6**	6
Total		80	3	17	100

The re-assessment did not seem to improve the overall level of agreement, which remained similar to that in the pilot study. While the observed proportion of agreement rose to 86%, the chance-corrected weighted kappa remained similar to the value obtained previously, at 0.7 (95% CI: 0.5–0.9). The marginal totals for the three categories remained symmetrically unbalanced with the category ‘current haematuria’ dominant for both raters. Therefore, the weighted kappa still tended to underestimate the true level of agreement.

Clarifying the definitions of each category did not affect the number of observations that were jointly classified as unclear, which remained at 2%.

In the re-assessment, raters 1 and 2 disagreed which category should be assigned to 14 observations in total (Table 8.3).

Table 8.3 Changes in areas of disagreement following clarification of category definitions

Source	Number of observations affected
Disagreement remained the same as per initial assessment	5
New disagreement	3
Raters both reversed their positions so that disagreement remained	6

8.1.4 Finalising the reference standard

Uninterpretable extracts (n=2) and extracts whose meaning the raters could not agree (n=14) were dropped. This left 84 extracts in the final reference standard, against which the performance of the classification system was assessed.

8.2 Performance of the final classification against the reference standard

The level of agreement between the final output of the classification process and the reference standard is reported in Table 8.4. As described in Section 6.7.3, this was done to validate the classification process and to provide a measure of its potential as a source of misclassification of the text string.

The observed proportion of agreement between the classification system and the reference standard was (75+6)/84 = 96%. The chance-corrected weighted Kappa was 0.9 (95% CI: 0.7–1.1).

Table 8.4 Two-way tabulation of the semi-automated classification procedure's output against that of the reference standard

		Reference standard		Total
		Current haematuria	**No current haematuria**	
Final classification output	Current haematuria	75	0	75
	Meaning unclear	3	0	3
	No current haematuria	0	6	6
Total		78	6	84

8.2.1 Sensitivity analyses

A sensitivity analysis was carried out as illustrated by the purple and black lines in Table 8.4. The category 'Meaning unclear' in the final classification output was merged first with 'Current haematuria' and then with 'No current haematuria' to derive the following binary systems:

1. ‘Current haematuria’ or “Not ‘Current haematuria’” (i.e. merged groups ‘No current haematuria’ and ‘Meaning unclear’). The black lines in Table 8.4 above illustrate this.

2. ‘No current haematuria’ or “Not ‘No current haematuria’” (i.e. merged groups ‘Current haematuria’ and ‘Meaning unclear’). The purple lines in Table 8.4 above illustrate this.

Two-way tabulation of the semi-automated classification procedure’s output against that of the gold standard is reported in Table 8.5, under sensitivity analysis 1.

Table 8.5 Two-way tabulation of the semi-automated classification procedure’s output against that of the reference standard under sensitivity analysis 1

		Reference standard		Totals
		Current haematuria	**No current haematuria**	
Final classification output	**Current haematuria**	75	0	75
	Not ‘Current haematuria’	3	6	9
Totals		78	6	84

In this analysis, the sensitivity is 75/78 (96%), the specificity and positive predictive value (PPV) are both 100%. The negative predictive value (NPV) is 6/9 (67%).

Two-way tabulation of the semi-automated classification procedure's output against that of the reference standard is reported in Table 8.6, under sensitivity analysis 2.

Under this analysis, sensitivity, specificity, PPV and NPV were all 100%.

Table 8.6 Two-way tabulation of the semi-automated classification procedure's output against that of the reference standard under sensitivity analysis 2

		Reference standard		Totals
		Current haematuria	**Not 'Current haematuria'**	
Final classification output	**Not 'No current haematuria'**	78	0	78
	No current haematuria	0	6	6
Totals		78	6	84

8.3 Other sources of uncertainty in text-based variables

Spelling and typographical errors made a minor contribution to uncertainty in text-based variables. The random sample contained 762 words, of which 5 were misspelt, representing an error rate of 0.7%. No instances of American instead of UK spelling – for example, 'anemia rather than 'anaemia' – were found.

8.4 Results of text extract processing

8.4.1 Raw data provided

The numbers of extracts provided by the CPRD are reported in Table 8.7. The search terms identified a small number of extracts that were inappropriate (e.g. 'denies sphincteric symptoms') – these were dropped from analysis (*n* = 9 in the pancreatic dataset; *n* = 8 in the bladder cancer dataset).

Table 8.7 Raw text data supplied by the CPRD

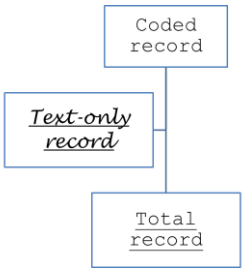
Cancer site	Source	No. of observations relating to:		
		Haematuria	Abdominal pain	Jaundice
Bladder	Clinical file	4,631	1,926	293
	Referral file	32	17	10
	Subtotal	*4,663*	*1,943*	*303*
	Extract repeated on separate date	8	6	1
	Total for classification	**4,671**	**1,949**	**304**
Pancreas	Clinical file	480	3,538	2,737
	Referral file	16	78	77
	Subtotal	*496*	*3,616*	*2,814*
	Extract repeated on separate date	0	2	1
	Total for classification	**496**	**3,618**	**2,815**

The CPRD provided a single copy of each unique extract. Identifying the date of recording (Step 1 – see Section 6.8.2) revealed that sometimes GPs duplicated the exact wording on separate dates. Replicate copies were created to ensure that the dataset contained the actual number of records made – one for each date – and the totals requiring classification are reported in Table 8.7.

8.4.2 Classification of text extracts

Reminder about variables

As a visual aid, fonts indicate the recording style used by the GP.



Courier New denotes variables generated from Read codes, i.e. haematuria, jaundice and abdominal pain.

Lucida Handwriting is used for variables generated from the text only, i.e. *haematuria, jaundice* and *abdominal pain.*

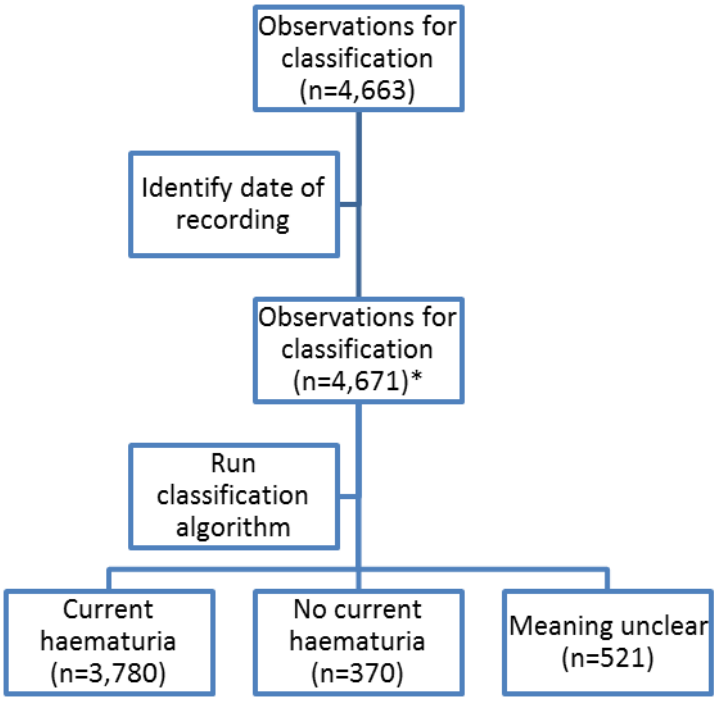
Underlined Courier New is used for the total record, i.e. where the variable is based on the Read code but is supplemented with records made solely in the free text, i.e. haematuria, jaundice and abdominal pain.

For clarity, the results of haematuria text extract classification in the bladder cancer dataset are illustrated in detail in Figure 8.1, Figure 8.2 and Figure 8.3. The results following classification of all the text extracts (i.e. for all three symptoms) are summarised in Table 8.8 and Table 8.13, respectively, for the bladder and pancreatic cancer datasets.

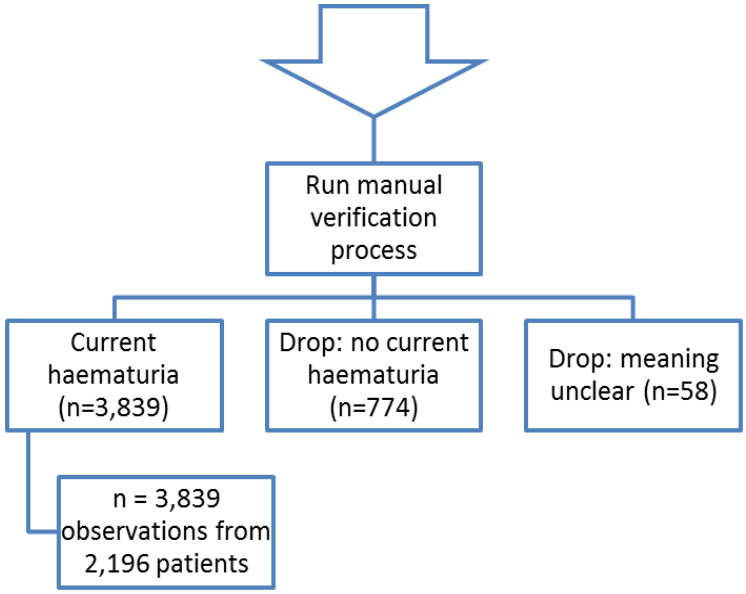
Figure 8.1 illustrates the results of initial classification of haematuria extracts by the algorithm and the numbers in each category after manual verification (see Section 6.8.2 for methods). The algorithm classified 3,780 observations as indicating current haematuria and 370 observations as 'no haematuria', leaving 521 whose meaning was unclear (Figure 8.1(a)).

Figure 8.1 Classification of text extracts: (a) by the algorithm and (b) after manual verification

(a)



(b)



*Only one copy of each unique text extract was supplied; however, GPs did enter exactly the same text on more than one occasion, so extracts were replicated to generate one copy for each date of recording.

After manual verification, the final classification identified 3,839 records (82.2%) of haematuria from 2,196 patients. 'No current haematuria' was assigned with certainty to 774 (16.6%) extracts, as well as to the 58 (1.2%) extracts whose meaning could not be determined (Figure 8.1(b)).

Figure 8.2 and Figure 8.3 illustrate the final steps of classification required to identify true text-only records of attendance for haematuria in the bladder cancer dataset (see Section 6.8.2 for the methods). Of the 3,839 attendances for haematuria identified from the text extracts, 2,699 were ‘hidden’ completely in the text as the GP had never recorded the attendance using a code.

Figure 8.2 Identifying the medcode paired with the text record (Step 1) and whether the paired medcode is related to the symptom discussed in the text (Step 2)

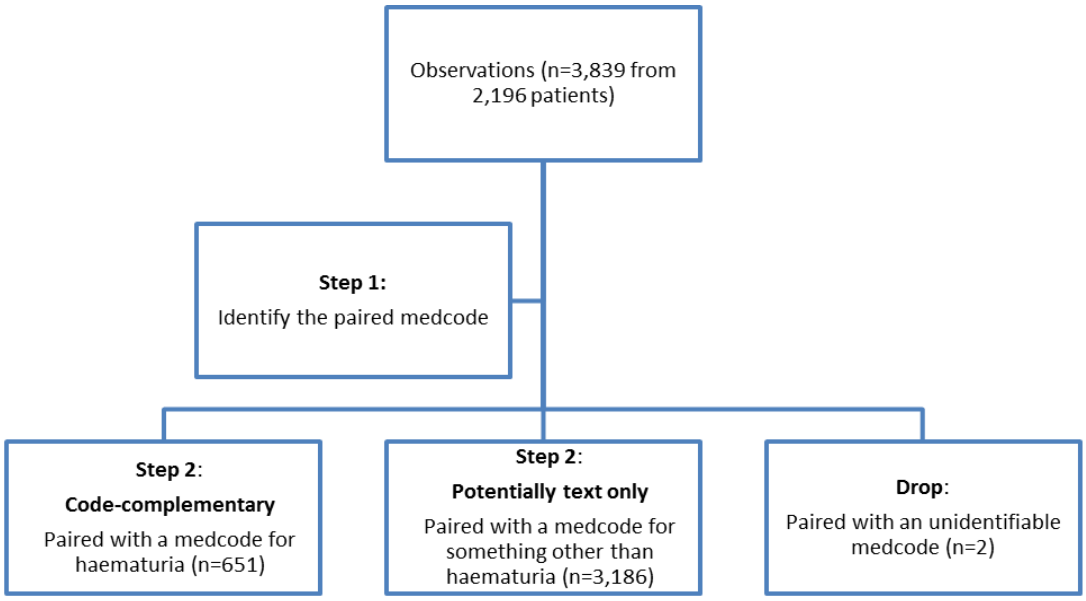
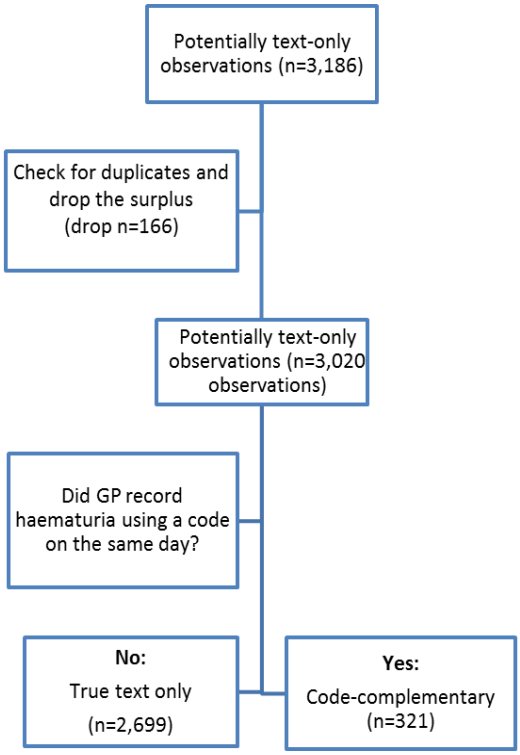


Figure 8.3 Step 3: Identifying true text-only recording



8.4.2.1 Classification of text extracts in the bladder cancer dataset

The details of classification of text extracts in the bladder cancer dataset are reported in Table 8.8. Potentially text-only records were identified for visible (*n*=3,188) and non-visible (*n*=311) haematuria, abdominal pain (*n*=1,208) and jaundice (*n*=105). The final numbers of true text-only records were 2,699 for visible and 298 for non-visible haematuria, 1,106 for abdominal pain and 88 for jaundice.

Table 8.8 Text extract classification for visible and non-visible haematuria, abdominal pain and jaundice in the bladder cancer dataset

Stage of classification		No. of observations in the bladder cancer dataset relating to:			
		Visible haematuria	Non-visible haematuria	Abdominal pain	Jaundice
By algorithm	Current symptom	3,780	314	1,356	149
	No current symptom	370	3,836	341	95
	Unclear	521	521	252	60
	Total	*4,671*	*4,671*	*1,949*	*304*
After manual check	**Current symptom**	**3,839**	**314**	**1,394**	**118**
	No current symptom	774	4,299	538	178
	Unclear	58	58	17	8
	Total	*4,671*	*4,671*	*1,949*	*304*

Stage of classification		No. of observations in the bladder cancer dataset relating to:			
		Visible haematuria	Non-visible haematuria	Abdominal pain	Jaundice
Step 2 [a]	Code-complementary [c]	651	3	186	13
	Potentially text only	**3,188**	**311**	**1,208**	**105**
	Total	*3,839*	*314*	*1,394*	*118*
Step 3 [b]	Dropped – paired medcode unidentifiable	2	0	0	0
	Dropped – duplicated record	166	10	42	8
	Code-complementary [c]	321	3	60	9
	True text-only record [d]	**2,699**	**298**	**1,106**	**88**
	Total	*3,188*	*311*	*1,208*	*105*

[a] Step 2: Identifying whether the paired medcode is related to the symptom discussed in the text (see Section 6.8.2).

[b] Step 3: Identifying true text-only recording (see Section 6.8.2).

[c] Code-complementary – the text extract was either paired with a medcode for the symptom, or the GP used a medcode to record the symptom on the same day elsewhere in the record.

[d] *Lucida Handwriting* is used for variables generated from the text only, i.e. *haematuria*, *jaundice* and *abdominal pain.*

8.4.2.2 Medcodes paired with true text-only records

The medcodes paired with the true text-only records were grouped into their appropriate Read code chapter, and are reported below for *visible* and *non-visible haematuria*, *abdominal pain* and *jaundice*, respectively.

8.4.2.2.1 Visible haematuria

For *visible haematuria*, approximately half of true text-only records were paired with general administrative codes such as 'Letter from specialist' and 'Telephone encounter' both in cases (1,264/2,462 = 51.3%) and controls (116/237 = 49.0%) (see Table 8.9). The next most frequent type of Read code in both patient groups was 'History/Symptoms', for example H/O: haematuria and dysuria. The text records paired with 'H/O: haematuria' were manually checked to see if they referred to a past rather than present occurrence of haematuria. If they did, this would suggest misclassification that required correction. In all cases the text suggested that haematuria was a current concern for the patient and that the classification was correct.

For 44 of the 2,462 (1.8%) attendances made by cases, visible haematuria was recorded in a text note paired with a diagnostic code for bladder cancer. These 44 attendances were made by 43 different cases (43/4,915 = 0.9%).

Table 8.9 Read code chapters associated with the medcode paired with the true text-only records of attendance for visible haematuria

Patient group	Read code chapter	Frequency	(%)
Controls	Administration	116	49.0
	History/Symptoms	40	16.9
	Genitourinary system disorders	27	11.4
	Other therapeutic procedures	15	6.3
	Preventive procedures	13	5.5
	Operations, procedures, sites	6	2.5
	Unspecified	5	2.1
	Examination/Signs	4	1.7
	[D]Symptoms	3	1.3
	Diagnostic procedures	2	0.8
	Laboratory procedures	2	0.8
	Endocrine, nutritional, metabolic and immunity disorders	1	0.4
	Circulatory system diseases	1	0.4
	Skin and subcutaneous tissue diseases	1	0.4
	Musculoskeletal and connective tissue diseases	1	0.4
	Subtotal	*237*	*100*
Cases	Administration	1,264	51.3
	History/Symptoms	345	14.0
	Other therapeutic procedures	218	8.9
	Preventive procedures	167	6.8
	Genitourinary system diseases	116	4.7
	Unspecified	84	3.4
	Operations, procedures, sites	66	2.7
	Neoplasms	44	1.8
	Laboratory procedures	36	1.5
	[D]Symptoms	33	1.3
	Examination/Signs	25	1.0
	Diseases of blood and blood-forming organs	17	0.7
	Diagnostic procedures	6	0.2
	Radiology physics	6	0.2
	Occupation	5	0.2

Patient group	Read code chapter	Frequency	(%)
	Circulatory system diseases	5	0.2
	Musculoskeletal and connective tissue diseases system	5	0.2
	Mental disorders	4	0.2
	Nervous system and sense organ diseases	3	0.1
	Respiratory system diseases	3	0.1
	Digestive system diseases	3	0.1
	Causes of injury poison	3	0.1
	Endocrine, nutritional, metabolic and immunity disorders	2	0.1
	Skin and subcutaneous tissue diseases	2	0.1
	Subtotal	*2,462*	*100*
Total		**2,699**	**100**

8.4.2.2.2 Non-visible haematuria

General administration was also the category of code most frequently paired with true text-only *non-visible haematuria* records; namely, in 30/50 (60%) of records in controls and in 116/248 (46.8%) of records in cases. The most frequently encountered general administrative codes were: 'Seen in urology clinic' (43/298 = 14.5%), 'Letter from specialist' (32/298 = 10.8%), 'Incoming mail NOS' (23/298 = 9.1%) and 'Patient reviewed' (20/298 = 6.7%). In the cases, codes related to genitourinary system diseases were paired with 36/248 (14.5%) of the text-only records, whereas in the controls this category was used in only 2/50 (4%) of text-only records (see Table 8.10).

Table 8.10 Read code chapters associated with the medcode paired with the true text-only records of attendance for non-visible haematuria

Patient group	Read code chapter	Frequency	%
Controls	Administration	30	60
	History/Symptoms	7	14
	Preventive procedures	3	6
	Laboratory procedures	2	4
	Operations, procedures, sites	2	4
	Genitourinary system diseases	2	4
	[D]Symptoms	2	4
	Other therapeutic procedures	1	2
	Unspecified	1	2
	Sub-total	*50*	*100*
Cases	Administration	116	46.8
	Genitourinary system diseases	36	14.5
	History/Symptoms	28	11.3
	Preventive procedures	21	8.5
	Other therapeutic procedures	15	6.1
	Laboratory procedures	11	4.4
	Unspecified	11	4.4
	Operations, procedures, sites	4	1.6
	[D]symptoms	2	0.8
	Diagnostic procedures	1	0.4
	Mental disorders	1	0.4
	Respiratory system diseases	1	0.4
	Digestive system diseases	1	0.4
	Sub-total	*248*	*100*
Total		**298**	**100**

Within the ‘History/Symptoms’ category, 41/298 (13.8%) of the text-only records were paired with codes for haematuria. These text records were manually checked to see if they had been misclassified, but were confirmed as reporting that non-visible haematuria was a current concern, but that visible haematuria had occurred in the past.

8.4.2.2.3 Abdominal pain

General administrative codes also accounted for the majority of medcodes paired with text-only *abdominal pain* records, both in cases (163/179= 42.9%) and controls (263/727= 36.2%) (Table 8.11). Indeed, the context of the consultation was important, with administrative codes such as 'Patient reviewed', 'Telephone encounter', and 'Home visit' accounting collectively for 219/1,106 (19.8%) of all text-only records. History/Symptoms was the next most frequent type of code, used for 69/379 (18.2%) of records in the cases and for 175/727 (24.1%) of records in the controls.

Table 8.11 Read code chapters associated with the medcode paired with the true text-only records of attendance for abdominal pain

Patient group	Read code chapter	Frequency	(%)
Controls	Administration	263	36.2
	History/Symptoms	175	24.1
	Preventive procedures	68	9.4
	Other therapeutic procedures	59	8.1
	Digestive system diseases	59	8.1
	Unspecified	40	5.5
	Genitourinary system diseases	10	1.4
	[D]Symptoms	9	1.2
	Examination/Signs	6	0.8
	Respiratory system diseases	6	0.8
	Infectious parasitic disease	5	0.7
	Musculoskeletal and connective tissue diseases	5	0.7
	Circulatory system diseases	4	0.6
	Radiology physics	3	0.4
	Operations, procedures, sites	3	0.4

Patient group	Read code chapter	Frequency	(%)
	Diseases of blood and blood-forming organs	3	0.4
	Causes of injury poison	3	0.4
	Endocrine, nutritional, metabolic and immunity disorders	2	0.3
	Occupation	1	0.1
	Mental disorders	1	0.1
	Nervous system and sense organ diseases	1	0.1
	Skin and subcutaneous disorders	1	0.1
	Sub-total	*727*	100
Cases	Administration	163	42.9
	History/Symptoms	69	18.2
	Other therapeutic procedures	37	9.7
	Genitourinary system diseases	32	8.4
	Preventive procedures	23	6.1
	Digestive system diseases	14	3.7
	Unspecified	13	3.7
	Examination/Signs	7	1.8
	[D]Symptoms	5	1.3
	Infectious parasitic disease	3	0.8
	Neoplasms	3	0.8
	Mental disorders	3	0.8
	Operations, procedures, sites	2	0.5
	Occupation	1	0.3
	Respiratory system diseases	1	0.3
	Skin and subcutaneous tissue diseases	1	0.3
	Musculoskeletal and connective tissue diseases	1	0.3
	Causes of injury poison	1	0.3
	Sub-total	*379*	100
Total		**1,106**	**100**

8.4.2.2.4 Jaundice

In the controls, Administration was the most frequent category of medcode paired with text-only *jaundice* records (24/26 = 38.7%), followed by

History/Symptoms (15/26 = 24.2%); however, for the cases, the two categories were used with similar frequency, i.e. 8/26 (30.8%) for History/Symptoms and 7/26 (26.9%) for Administration (Table 8.12).

Table 8.12 Read code chapters associated with the medcode paired with the true text-only records of attendance for jaundice

Patient group	Read code chapter	Frequency	%
Controls	Administration	24	38.7
	History/Symptoms	15	24.2
	Preventive procedures	4	6.5
	Other therapeutic procedures	4	6.5
	Unspecified	4	6.5
	Operations, procedures, sites	3	4.8
	[D]Symptoms	3	4.8
	Circulatory system diseases	2	3.2
	Digestive system diseases	2	3.2
	Neoplasms	1	1.6
	Sub-total	*26*	*100*
Cases	History/Symptoms	8	30.8
	Administration	7	26.9
	Preventive procedures	2	7.7
	Neoplasms	2	7.7
	Genitourinary system diseases	2	7.7
	Laboratory procedures	1	3.9
	Operations, procedures, sites	1	3.9
	Other therapeutic procedures	1	3.9
	Respiratory system diseases	1	3.9
	Digestive system diseases	1	3.9
	Sub-total	*26*	*100*
Total		**88**	100

8.4.2.3 Classification of text extracts in the pancreatic cancer dataset

The details of classification of text extracts in the pancreatic cancer dataset are reported in Table 8.13.

Table 8.13 Text extract classification for jaundice, abdominal pain and haematuria in the pancreatic cancer dataset

Stage of classification		No. of observations in the pancreatic cancer dataset relating to:		
		Jaundice	Abdominal pain	Visible haematuria
Classification by algorithm	**Current symptom**	**2,391**	**2,852**	**318**
	No current symptom	285	395	77
	Unclear	139	371	101
	Total	*2,815*	*3,618*	*496*
Classification after manual verification	**Current symptom**	**2,243**	**2,979**	**308**
	No current symptom	432	573	156
	Unclear	140	66	32
	Total	*2,815*	*3,618*	*496*
Step 2 [a]	Code-complementary [c]	308	458	39
	Potentially text only	**1,935**	**2,521**	**269**
	Total	*2,243*	*2,979*	*308*
Step 3 [b]	Dropped (paired medcode unidentifiable)	0	0	0
	Dropped (duplicated record)	133	102	8
	Code-complementary [c]	163	204	10
	True text-only record	**1,639**	**2,215**	**251**
	Total	1,935	*2,521*	*269*

[a] Step 2: Identifying whether the paired medcode is related to the symptom discussed in the text (see Section 6.8.2).

[b] Step 3: Identifying true text-only recording (see Section 6.8.2).

[c] Code-complementary – the text extract was either paired with a medcode for the symptom, or the GP used a medcode to record the symptom on the same day elsewhere in the record.

Potentially text-only records were identified for jaundice (*n* = 1,935), abdominal pain (*n* = 2,521) and visible haematuria (*n* = 269). The final numbers of true text-only records for *jaundice*, *abdominal pain* and *visible haematuria* were 1,639, 2,215 and 251, respectively.

8.4.2.4 Medcodes paired with text-only records

The medcodes paired with the true text-only records were grouped into their appropriate Read code chapter, and are reported below for *jaundice*, *abdominal pain* and *visible haematuria*, respectively.

8.4.2.4.1 Jaundice

Text-only *jaundice* records were most frequently paired with general administrative codes both in controls (19/35 = 54.3%) and cases (720/1,604 = 44.9%). The second most common category of code was History/Symptoms, used for 7/35 (20.0%) of records in controls and 233/1,604 (14.5%) in cases (Table 8.14). This included 40 text-only *jaundice* records paired with codes indicating that the patient had a history of jaundice. The individual text snippets were re-checked and it was confirmed that they all indicated that jaundice was a current clinical concern for these patients, i.e. that the classification was correct. In the controls the remaining text-only records were paired with codes related to

the history, examination or other procedures, whereas in the cases codes for specific diseases and diagnoses tended to be used. For 83/1,604 (5.2%) of the attendances made by cases, jaundice was noted in a text record paired with a diagnostic code for pancreatic cancer. These 83 attendances were made by 77 of the 3,635 (2.1%) cases.

Table 8.14 Read code chapters associated with the medcode paired with the true text-only records of attendance for jaundice

Patient group	Read code chapter	Frequency	%
Controls	Administration	19	54.3
	History/Symptoms	7	20.0
	Other therapeutic procedures	3	8.6
	Preventive procedures	2	5.7
	[D]Symptoms	2	5.7
	Examination/Signs	1	2.9
	Laboratory procedures	1	2.9
	Sub-total	*35*	*100*
Cases	Administration	720	44.9
	History/Symptoms	233	14.5
	Other therapeutic procedures	162	10.1
	Preventive procedures	138	8.6
	Neoplasms	83	5.2
	Examination/Signs	57	3.6
	[D]Symptoms	42	2.6
	Unspecified	41	2.6
	Operations procedures sites	37	2.3
	Skin and subcutaneous tissue diseases	36	2.2
	Digestive system diseases	19	1.2
	Laboratory procedures	13	0.8
	Diagnostic procedures	4	0.3
	Radiology physics	3	0.2
	Infectious parasitic disease	3	0.2
	Genitourinary system diseases	3	0.2
	Endocrine, nutritional, metabolic and immunity disorders	2	0.1
	Circulatory system diseases	2	0.1
	Musculoskeletal and connective tissue	2	0.1

Patient group	Read code chapter	Frequency	%
	diseases		
	Occupation	1	0.06
	Mental disorders	1	0.06
	Respiratory system diseases	1	0.06
	Causes of injury poison	1	0.06
	Sub-total	*1,604*	*100*
Total		**1,639**	**100**

8.4.2.4.2 Abdominal pain

Text-only *abdominal pain* records were most frequently paired with administration codes (226/594 = 38.1% of records in controls, and 705/1,621 = 43.5% of records in cases), followed once again by codes for History/Symptoms (controls: 147/594 = 24.8%; cases: 289/1,621 = 17.8%) (see Table 8.15).

Table 8.15 Read code chapters associated with the medcode paired with the true text-only records of attendance for abdominal pain

Patient group	Read code chapter	Frequency	%
Controls	Administration	226	38.1
	History/Symptoms	147	24.8
	Other therapeutic procedures	52	8.8
	Digestive system diseases	45	7.6
	Preventive procedures	34	5.7
	[D]Symptoms	14	2.4
	Genitourinary system diseases	13	2.2
	Examination/Signs	12	2.0
	Unspecified	12	2.0
	Musculoskeletal and connective tissue diseases	7	1.2
	Laboratory procedures	6	1.0
	Operations, procedures, sites	6	1.0
	Infectious parasitic disease	6	1.0
	Causes of injury poison	3	0.5
	Diseases of blood and blood-forming organs	2	0.3

Patient group	Read code chapter	Frequency	%
	Mental disorders	2	0.3
	Injury poison	2	0.3
	Neoplasms	1	0.2
	Nervous system and sense organ diseases	1	0.2
	Circulatory system diseases	1	0.2
	Respiratory system diseases	1	0.2
	Skin and subcutaneous tissue diseases	1	0.2
	Sub-total	*594*	*100*
Cases	Administration	705	43.5
	History/Symptoms	289	17.8
	Preventive procedures	188	11.6
	Other therapeutic procedures	150	9.3
	Unspecified	59	3.6
	Digestive system diseases	56	3.5
	[D]Symptoms	47	2.9
	Examination/Signs	43	2.7
	Neoplasms	18	1.1
	Genitourinary system diseases	11	0.7
	Laboratory procedures	10	0.6
	Musculoskeletal and connective tissue diseases	10	0.6
	Operations, procedures, sites	7	0.4
	Circulatory system diseases	5	0.3
	Occupation	4	0.3
	Skin and subcutaneous tissue diseases	4	0.3
	Radiology physics	3	0.2
	Endocrine, nutritional, metabolic and immunity disorders	3	0.2
	Respiratory system diseases	2	0.1
	Diagnostic procedures	1	0.06
	Infectious parasitic disease	1	0.06
	Diseases of blood and blood-forming organs	1	0.06
	Mental	1	0.06
	Nervous system and sense organ diseases	1	0.06
	Injury poison	1	0.06
	Causes of injury poison	1	0.06
	Sub-total	*1,621*	*100*
Total		**2,215**	**100**

8.4.2.4.3 Visible haematuria

Text-only *visible haematuria* records were most frequently paired with medcodes from the administration category (controls: 68/151 = 44.7%; cases: 43/100 = 43%), followed by History/Symptom codes (controls: 23/151 = 15.1%; cases: 19/100 = 19%).

Table 8.16 Read code chapters associated with the medcode paired with the true text-only records of attendance for visible haematuria

Patient group	Read code chapter	Frequency	%
Controls	Administration	68	44.7
	History/Symptoms	23	15.1
	Other therapeutic procedures	17	11.2
	Genitourinary system diseases	17	11.2
	Preventive procedures	7	4.6
	[D]Symptoms	5	3.3
	Operations, procedures, sites	4	2.6
	Examination/Signs	2	1.3
	Laboratory procedures	2	1.3
	Circulatory system diseases	2	1.3
	Unspecified	1	0.7
	Diagnostic procedures	1	0.7
	Mental disorders	1	0.7
	Causes of injury poison	1	0.7
	Sub-total	*151*	*100*
Cases	Administration	43	43
	History/Symptoms	19	19
	Other therapeutic procedures	9	9
	Preventive procedures	8	8
	Genitourinary system diseases	8	8
	[D]Symptoms	4	4
	Neoplasms	2	2
	Unspecified	2	2
	Occupation	1	1
	Diagnostic procedures	1	1
	Laboratory procedures	1	1
	Diseases of blood and blood-forming organs	1	1
	Digestive system diseases	1	1
	Sub-total	*100*	*100*
Total		**251**	100

9 Results: Recording style

9.1 Event-level data

In this analysis the aim was to identify the recording style of attendances for the symptoms. The analysis addresses research questions 1: ‘How much symptom information is documented in electronic medical records using text rather than a code?’ and 3: ‘Does recording style vary with type of symptom?’ (Section 5).

The total numbers of attendances for each of haematuria, abdominal pain and jaundice were obtained by adding the number of text-only records of attendance to the numbers of coded records.

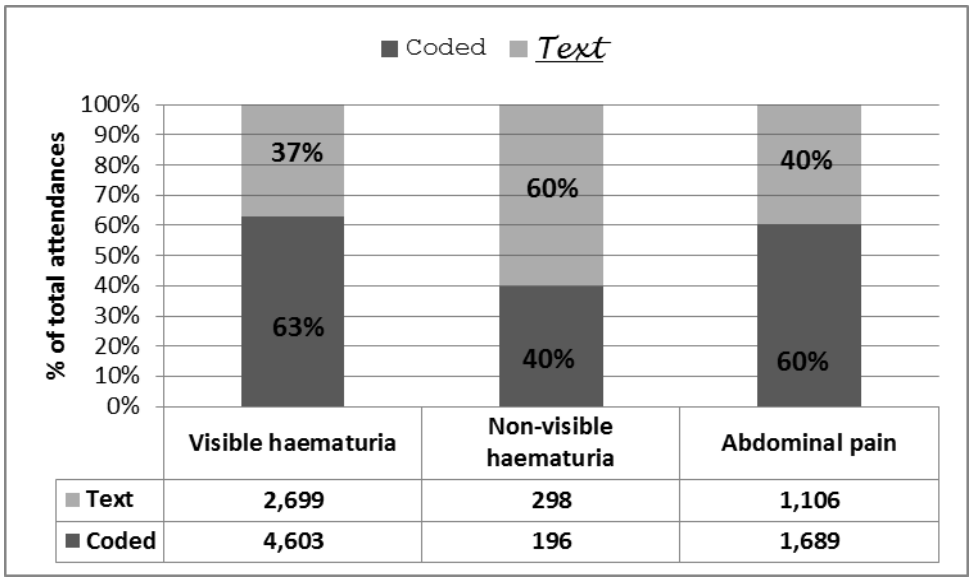
9.1.1 Recording style at the event level in the bladder cancer dataset

The method used to record attendances for visible and non-visible haematuria and abdominal pain is reported in Figure 9.1. The numbers of attendances recorded in code (i.e. `visible` and `non-visible haematuria` and `abdominal pain`) and solely in the text (i.e. *visible* and *non-visible haematuria* and *abdominal pain*) are tabulated at the bottom of Figure 9.1. The percentage of the total number of attendances obtained from coded and from text-only records is shown graphically at the top of the figure.

Note these figures were obtained by re-using the symptom thesauri created for the original studies; therefore, ‘abdominal pain’ includes codes for abdominal

tenderness, rigidity and cramps, but not for epigastric pain, indigestion or dyspepsia (see Appendix 4: Symptom thesauri).

Figure 9.1 Event-level data: method used to record attendances for visible and non-visible haematuria and abdominal pain in the bladder cancer dataset



	Visible haematuria	Non-visible haematuria	Abdominal pain
Text	2,699	298	1,106
Coded	4,603	196	1,689

Note: The percentages of the total number of attendances with a symptom are marked on the bars, and the raw numbers are tabulated beneath.

As shown in Figure 9.1, while large numbers of text-only *visible haematuria* and *abdominal pain* records were made in the bladder cancer dataset, coding was the favoured method. Indeed, coded `visible haematuria` (4,603/7,302=63%) and `abdominal pain` (1,689/2,795=60%) records accounted for nearly two-thirds of the overall numbers of attendances for these symptoms. Lower numbers of attendance were observed for non-visible haematuria compared with both visible haematuria and abdominal pain. However, text-only was the preferred recording style, with *non-visible*

haematuria records accounting for nearly two-thirds of all the attendances for this feature (298/494=60%).

9.1.2 Recording style at the event level in the pancreatic cancer dataset

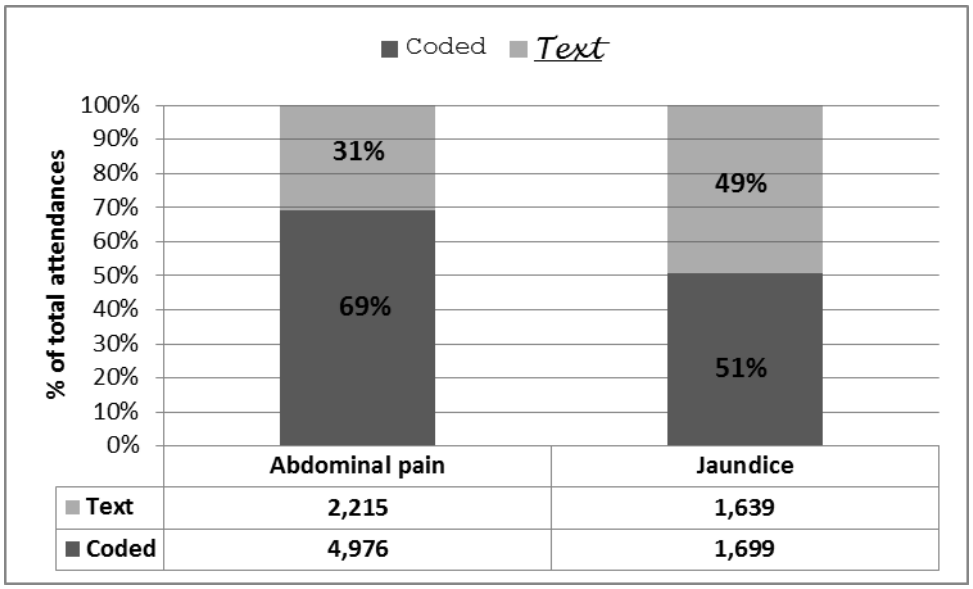
The method used to record attendances for jaundice and abdominal pain is reported in Figure 9.2. The numbers of attendances recorded in code (i.e. `jaundice` and `abdominal pain`) and solely in the text (i.e. *jaundice* and *abdominal pain*) are tabulated at the bottom of Figure 9.2. The percentage of the total number of attendances obtained from coded and from text-only records is shown graphically at the top of the figure.

For jaundice there were similar numbers of `jaundice` (1,699/3,338 = 51%) and *jaundice* (1,639/3,338 = 49%) records, such that there was no overall preference for either recording style (Figure 9.2).

As shown in Figure 9.2, large numbers of text-only *abdominal pain* records were made in the pancreatic cancer dataset. Despite this, coding was the preferred method, with `abdominal pain` records accounting for more than two-thirds (4,976/7,191 = 69%) of the attendances for this symptom.

The coded `abdominal pain` records were obtained re-using the symptom thesaurus created in the original pancreatic cancer study. The thesaurus contained codes for epigastric pain, indigestion and dyspepsia, but not abdominal tenderness, rigidity or cramps (see Appendix 4: Symptom thesauri).

Figure 9.2 Event-level data: method used to record attendances for jaundice and abdominal pain in the pancreatic cancer dataset



Note: The percentages of the total number of attendances with a symptom are marked on the bars, and the raw numbers are tabulated beneath.

9.1.3 Medcodes used to record symptoms

The symptom thesauri were deliberately comprehensive, in order to maximise the capture of all coded records of symptoms. The results below indicate how many of the medcodes included in the thesauri were actually used by the GPs.

9.1.3.1 Visible and non-visible haematuria – bladder cancer

Of the 26 medcodes included in the visible haematuria symptom thesaurus, 12 were never selected by GPs in the bladder cancer dataset. In fact, the majority of coded `visible haematuria` records were made using a single medcode (507: 'haematuria'), which accounted for 3,769 of the 4,603 (81.9%)

attendances in the analysis period (see Table 9.1). Six more medcodes were used to identify a further 18% of the coded records, leaving just a handful of matches for the remaining six medcodes that were ever selected.

Table 9.1 Event-level data: codes used to document patient attendances for visible haematuria in the bladder cancer dataset

Medcode	Description	Frequency	(%)
507	Haematuria	3,769	81.9
7232	Frank haematuria	271	5.9
6030	Haematuria - symptom	206	4.5
6659	Blood in urine - haematuria	135	2.9
9651	Painless haematuria	122	2.6
6234	Blood in urine - symptom	66	1.4
17060	Recurrent and persistent haematuria	24	0.5
6901	Clot haematuria	4	0.1
20357	Painful haematuria	2	0.04
13915	RBCs- red blood cells in urine	1	0.02
13929	Urine blood test = +	1	0.02
13934	Urine blood test = ++	1	0.02
19792	Urine blood test = +++	1	0.02
Total		**4,603**	**100**

Only one medcode for non-visible haematuria (277: microscopic haematuria) had been identified and included in the thesaurus for this symptom; therefore, this accounted for 100% of the attendances.

9.1.3.2 Jaundice – pancreatic cancer

A small number of medcodes were used to document attendances for jaundice (Table 9.2). Indeed, just four medcodes identified 1,479 of the total 1,699 (87.1%) attendances for this symptom. Only one medcode included in the symptom thesaurus was never used, i.e. [D]Jaundice (not of newborn) NOS.

Table 9.2 Event-level data: codes used to document patient attendances for jaundice in the pancreatic cancer dataset

Medcode	Description	Frequency	(%)
3121	Obstructive jaundice NOS	636	37.4
6000	Jaundice - symptom	426	25.1
2612	[D]Jaundice	210	12.4
355	[D]Jaundice (not of newborn)	207	12.2
5996	O/E - jaundiced	156	9.2
25418	Yellow/jaundiced colour	43	2.5
29488	O/E - jaundiced colour	9	0.5
18019	Yellow - symptom	6	0.4
18574	[D]Icterus NOS	6	0.4
Total		**1,699**	**100**

9.1.3.3 Abdominal pain

As noted above, the original bladder and pancreatic cancer studies differed in their choice of codes used to identify attendances for abdominal pain (see

Appendix 4: Symptom thesauri). This was largely because, in the pancreatic cancer study, abdominal pain, dyspepsia and indigestion were all combined into a composite 'abdominal' symptom. In addition, pain from the bladder refers to the hypogastrium (aka suprapubic region), while pain from the pancreas refers to the left hypochondrium and epigastric region (see Section 4.5.3.4.4).

The frequency of use of medcodes is reported separately for each dataset below.

9.1.3.3.1 Bladder cancer dataset

Two specific medcodes (177: abdominal pain and 1763: [D]abdominal pain) accounted for 1,244 of the total 1,689 (73.7%) attendances for abdominal pain in the bladder cancer dataset (Table 9.3).

Table 9.3 Event-level data: codes used to document patient attendances for abdominal pain in the bladder cancer dataset

Medcode	Description	Frequency	(%)
177	Abdominal pain	815	48.3
1763	[D]Abdominal pain	429	25.4
2383	Abdominal discomfort	139	8.2
1976	Abdominal pain type	122	7.2
7812	Colicky abdominal pain	27	1.6
2056	[D]Abdominal colic	25	1.5
22608	Lower abdominal pain	23	1.4
5782	O/E - abdomen tender	22	1.3

Medcode	Description	Frequency	(%)
5960	Site of abdominal pain	17	1.0
3338	Central abdominal pain	14	0.8
4617	[D]Abdominal pain NOS	14	0.8
716	[D]Abdominal cramps	12	0.7
7726	[D]Right upper quadrant pain	6	0.4
8436	[D]Upper abdominal pain	4	0.2
3978	Right upper quadrant pain	3	0.2
4771	Upper abdominal pain	3	0.2
8362	[D]Left upper quadrant pain	3	0.2
9695	[D]Nonspecific abdominal pain	3	0.2
19283	[D]Umbilical pain	3	0.2
2234	General abdominal pain-symptom	1	0.06
9061	Generalised abdominal pain	1	0.06
11070	[D]Abdominal tenderness	1	0.06
16402	[D]Left lower quadrant pain	1	0.06
24661	[D]Recurrent acute abdominal pain	1	0.06
Total		**1,689**	**100**

9.1.3.3.2 Pancreatic cancer dataset

As with the bladder cancer dataset, a small number of medcodes accounted for the majority (4,168/4,976=83.8%) of the attendances for ‘abdominal pain’ in the analysis period (see Table 9.4). These codes were: 177 (abdominal pain), 257 (dyspepsia), 1763 ([D]abdominal pain) and 290 (epigastric pain).

Table 9.4 Event-level data: codes used to document patient attendances for abdominal pain in the pancreatic cancer dataset

Medcode	Description	Frequency	(%)
177	Abdominal pain	1,670	33. 6
257	Dyspepsia	889	17.9
1763	[D]Abdominal pain	835	16.8
290	Epigastric pain	774	15.5
1976	Abdominal pain type	273	5.5
134	Indigestion	178	3.6
5862	Indigestion symptoms	136	2.7
2056	[D]Abdominal colic	50	1.0
3338	[D]Abdominal pain NOS	37	0.7
5960	Site of abdominal pain	32	0.6
4617	Central abdominal pain	26	0.5
542	[D]Epigastric pain	18	0.4
15180	O/E - abdo. pain on palpation	11	0.2
19283	[D]Nonspecific abdominal pain	10	0.2
7623	Indigestion NOS	8	0.2
43233	Undiagnosed dyspepsia	7	0.1
24661	Generalised abdominal pain	6	0.1
8697	Flatulent dyspepsia	4	0.08
20640	O/E - epigastric pain on palp.	4	0.08
14916	Indigestion symptom NOS	2	0.04
37118	O/E -abd.pain on palpation NOS	2	0.04
2234	O/E - abd. pain - epigastrium	1	0.02

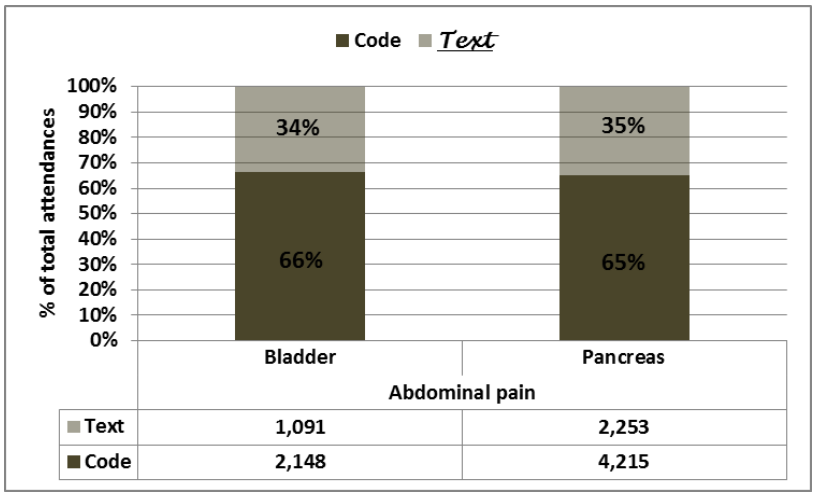
Medcode	Description	Frequency	(%)
9061	Type of GIT pain NOS	1	0.02
19223	[D]Left lower quadrant pain	1	0.02
44484	[D]Recurrent acute abdominal pain	1	0.02
Total	Total	**4,976**	**100**

9.1.3.3.3 Using a generic abdominal pain thesaurus for both cancer sites

The results in Figure 9.1 and Figure 9.2 suggested that the preference for coding abdominal pain was greater in the pancreatic (4,976/7,191 = 69%) than in the bladder (1,689/2,795 = 60%) cancer dataset.

In separate analysis, a generic symptom thesaurus for abdominal pain was used to allow direct comparison of recording patterns between the two cancer datasets (Figure 9.3).

Figure 9.3 Event-level data: abdominal pain recording, with coded records identified using the same symptom thesaurus in both cancer sites



In this analysis, the preference for coding was similar in the two datasets; namely, 2,148/3,239 (66%) in bladder, and 4,215/6,468 (65%) in pancreas.

9.2 Patient-level data

Event-level data (reported above) equate to the number of *attendances* for each symptom within the analysis period (including multiple attendances per patient). In contrast, patient-level data report the numbers of patients (*attendees*) attending at least once for the symptom of interest.

The results in this section relate to research questions 1: 'How much symptom information is documented in electronic medical records using text rather than a code?'; 2: Are studies of coded data vulnerable to bias arising from the differential use of text and codes between comparison groups'; 3: 'Does recording style vary with type of symptom?'; and 4: 'Does the recording style vary with the clinical context of presentation of a symptom' (see Section 5).

9.2.1 Overall recording style preference

The recording style at the patient level[a] for alarm and non-alarm symptoms in the bladder and pancreatic cancer datasets is reported in Table 9.5. The numbers of patients attending at least once because of the symptom in question

[a] Each patient was categorised by the overall style used to record their attendances for the symptom. 'Coded' was assigned if *any* record of a symptom was in coded form; conversely, 'text-only' was designated only when *all* instances were noted solely in the text.

were obtained using the bespoke symptom thesauri created for the original studies.

The results suggest that the original studies omitted at least 20% and as many as 59% of patients who had attended with haematuria (visible or non-visible), jaundice or abdominal pain, justifying our concerns about omitting text records from analysis.

In the bladder cancer dataset, coded `visible haematuria` (2,787/3,483 = 80%) and `abdominal pain` (1,142/1,762 = 65%) records identified a greater proportion of attendees for these symptoms compared with text-only *visible haematuria* (696/3,483 = 20%) and *abdominal pain* (620/1,762 = 35%) records (Figure 9.4 and Table 9.5).

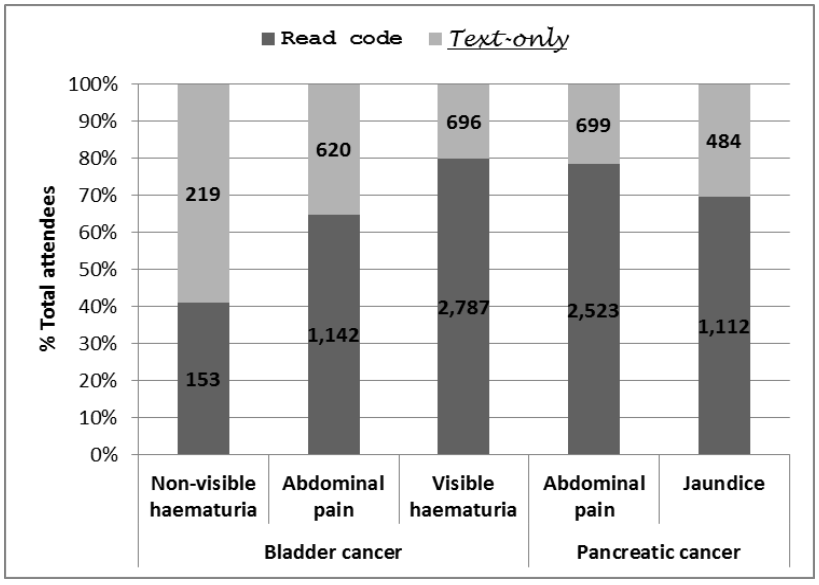
In contrast, text-only *non-visible haematuria* records (219/372 = 59%) identified more attendees for this symptom compared with coded `non-visible haematuria` records (153/372 = 41%) (see Figure 9.4).

In the pancreatic cancer dataset, coded records for both `jaundice` (1,112/1,596 = 70%) and `abdominal pain` (2,523/3,222 = 78%) identified a greater proportion of attendees for these symptoms compared with the text-only *jaundice* (484/1,596 = 30%) and *abdominal pain* (699/3,222 = 22%) records (see Figure 9.4).

Table 9.5 Symptom recording style for alarm symptoms (visible haematuria in bladder cancer; jaundice in pancreatic cancer) and non-alarm symptoms (abdominal pain for both cancers; non-visible haematuria for bladder cancer), regardless of patient status

Cancer site	Feature	Number (% of total) of patients attending at least once where the event was recorded using:		
		Read code, *n* (%)	*Text-only*, *n* (%)	**Total**, *n* (%)
Bladder	Non-visible haematuria	153 (41)	219 (59)	372 (100)
	Abdominal pain	1,142 (65)	620 (35)	1,762 (100)
	Visible haematuria	2,787 (80)	696 (20)	3,483 (100)
Pancreas	Abdominal pain	2,523 (78)	699 (22)	3,222 (100)
	Jaundice	1,112 (70)	484 (30)	1,596 (100)

Figure 9.4 Graphical presentation of the recording style for symptoms within bladder and pancreatic cancer



9.2.2 Association between symptom recording style and patient factors

In this analysis, the aim was to compare recording styles between cases and controls to address research question 2: ‘Are studies of coded data vulnerable to bias arising from the differential use of text and codes between comparison groups?’ (see Section 5).

A second aim was to examine whether associations between recording style and patient status varied with gender, to address research question 4: ‘Does the recording style vary with the clinical context of presentation of a symptom?’

9.2.2.1 In the bladder cancer dataset

The recording styles that identified attendees for visible and non-visible haematuria and for abdominal pain are plotted in Figure 9.5, Figure 9.8 and Figure 9.10, respectively. The raw data are reported in Table 9.6.

Table 9.6 Numbers of patients attending at least once in the analysis period with abdominal pain or haematuria (visible or non-visible) grouped by recording style

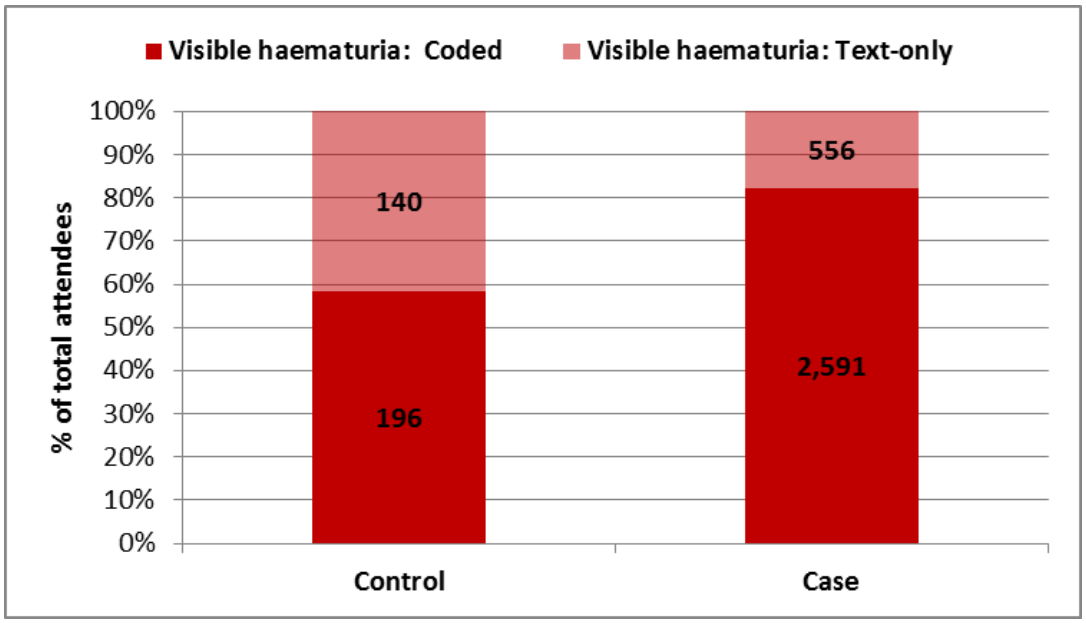
Clinical feature and its recording style [a]	Number of cases (*n* = 4,915) and controls (*n* = 21,718) attending at least once in the analysis period for the feature		Total, *n* (% of total with the symptom)
	Cases, *n* (% of cases)	Controls, *n* (% of controls)	
Visible haematuria	2,591 (52.7)	196 (0.9)	2,787 (80.0)
Visible haematuria	556 (11.3)	140 (0.6)	696 (20.0)
Total: Visible haematuria	3,147 (64.0)	336 (1.5)	3,483 (100)
Non-visible haematuria	127 (2.6)	26 (0.1)	153 (41.1)
Non-visible haematuria	185 (3.8)	34 (0.2)	219 (58.9)
Total: Non-visible haematuria	312 (6.4)	60 (0.3)	372 (100)
Abdominal pain	358 (7.3)	784 (3.6)	1,142 (64.8)
Abdominal pain	189 (3.8)	431 (2.0)	620 (35.2)
Total: Abdominal pain	547 (11.1)	1,215 (5.6)	1,762 (100)

[a] Reminder that fonts are used to indicate recording style; for example, Visible haematuria – GP used a Read code for some or all attendances; *Visible haematuria* – GPs used text-only to record all attendances; Visible haematuria – total record, i.e. codes plus text-only records.

9.2.2.1.1 Visible haematuria

There was strong evidence of an association between patient status (case or control) and the recording style that identified attendees for visible haematuria (χ2 test, p<0.0001). Coded `visible haematuria` records identified a greater proportion of the cases (2,591/3,147 = 82%) than of the controls (196/336 = 58%) who had attended for visible haematuria during the analysis period.

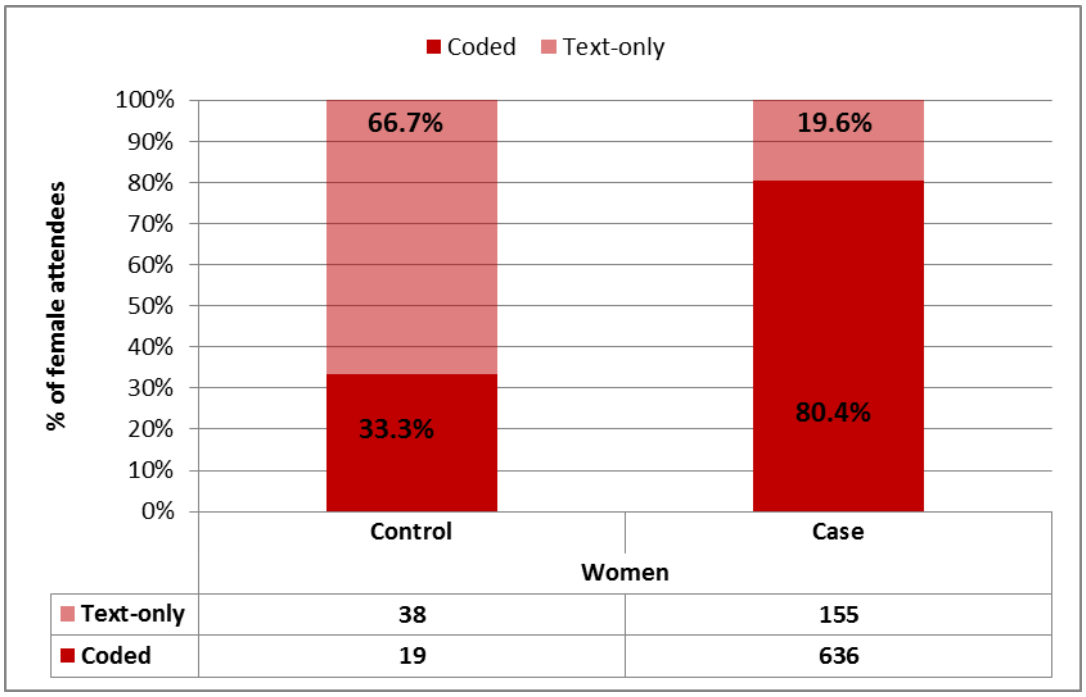
Figure 9.5 The number of patients who attended at least once in the analysis period for visible haematuria where the event was recorded using a code (dark red) or text-only (light red) in bladder cancer cases (*n* = 3,147/4,915) and controls (*n* = 336/21,718) separately



This association between recording style and case/control status was markedly different in women than in men, as shown in Figure 9.6 and Figure 9.7, respectively.

In women, there was a strong association between recording style and case/control status (χ2 test, p<0.0001) (Figure 9.6). Coded records identified 636 of the 791 (80.4%) female cases who had experienced at least one episode of visible haematuria – a proportion similar to that observed for the cases overall. In contrast, coded records identified only 19 of the 57 (33.3%) female controls who ever attended for this symptom.

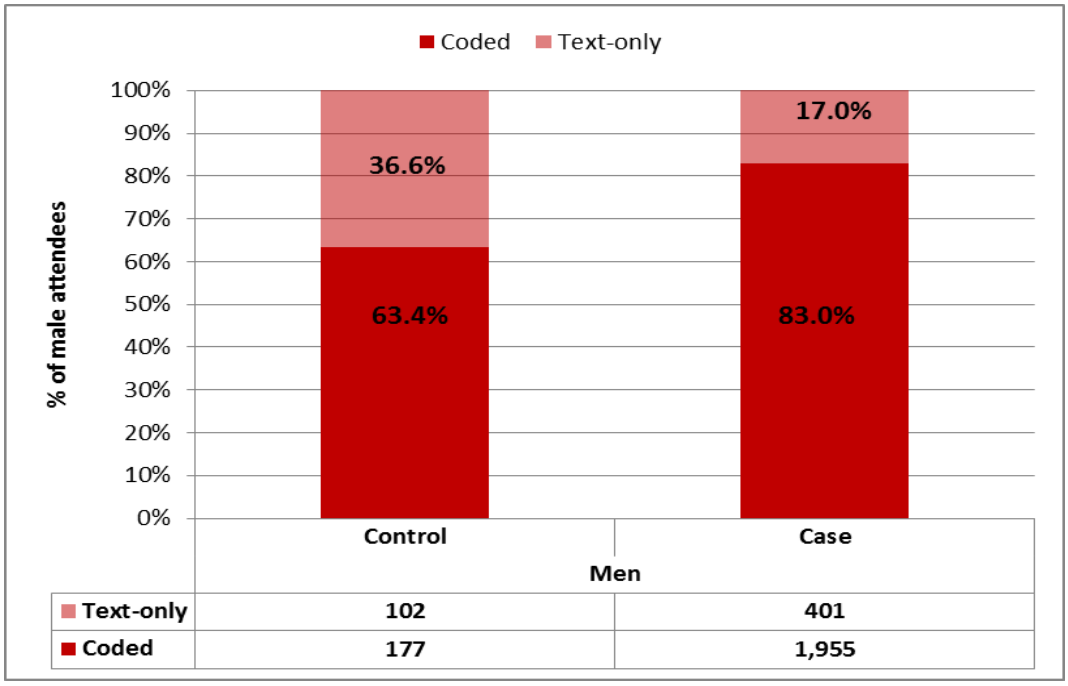
Figure 9.6 The number of women who attended at least once in the analysis period for visible haematuria where the event was recorded using a code (dark red) or text-only (light red) in bladder cancer cases (*n* = 791/1,352) and controls (*n* = 57/6,266) separately



There was also a strong association between recording style and case/control status in men (χ2 test, p<0.0001). Compared with the female cases, coded records identified a similar proportion of male cases who attended at least once for visible haematuria (1,955/2,356 = 83.0%). For the controls, there was a

marked difference in recording style between men and women. Coded records identified 177 of the 279 (63.4%) male controls with visible haematuria. In other words, the extent of data hidden in text-only records for male controls was considerably less than that for the female controls.

Figure 9.7 The number of men who attended at least once in the analysis period for visible haematuria where the event was recorded using a code (dark red) or text-only (light red) in bladder cancer cases (*n* = 2,356/3,563) and controls (*n* = 57/6,266) separately

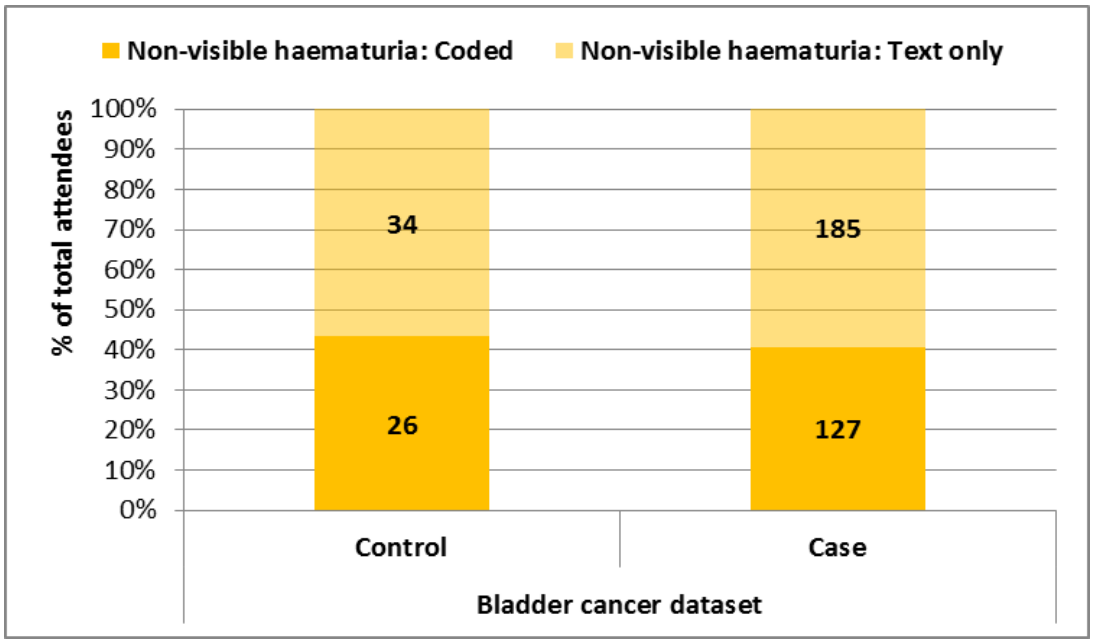


9.2.2.1.2 Non-visible haematuria

In contrast with the visible form of haematuria, there was no evidence of an association between patient (men and women combined) status and recording style used to identify attendees for non-visible haematuria ($\chi 2$ test, $p = 0.7$). Text-only *non-visible haematuria* records predominated, identifying similar

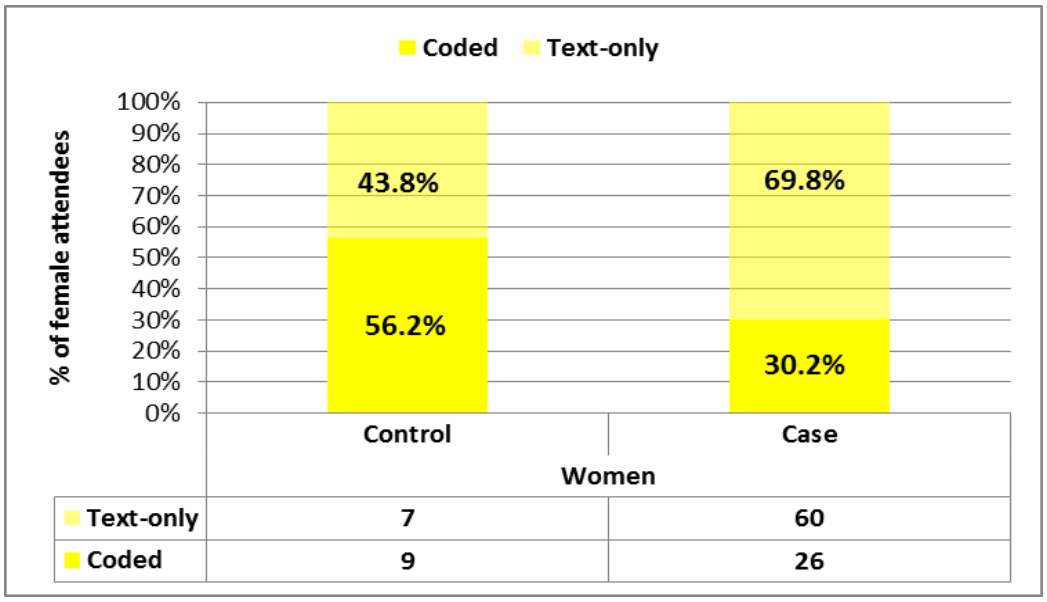
proportions of cases (185/312 = 59%) and controls (34/60 = 57%) who had attended for non-visible haematuria during the analysis period.

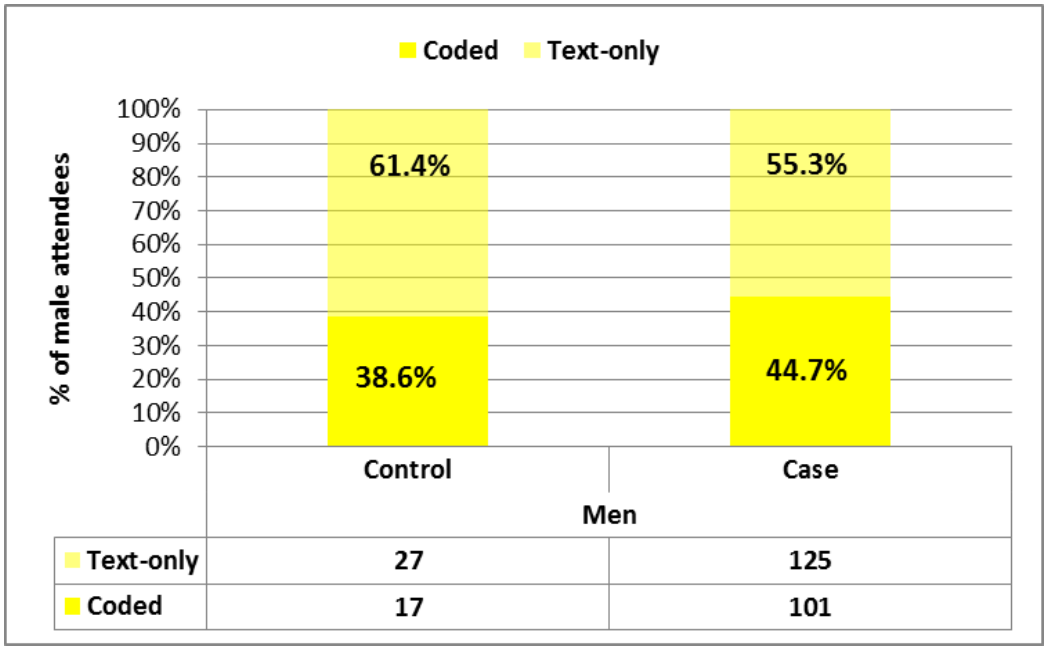
Figure 9.8 The number of patients who attended at least once in the analysis period for non-visible haematuria where the event was recorded using Read codes (dark yellow) and text-only (light yellow) in bladder cancer cases (*n* = 312/4,915) and controls (*n* = 60/21,718) separately



Examining the data for men and women separately indicated an association between recording style and case/control status in the women ($p<0.05$) but not the men ($p = 0.5$), although the numbers were small (see Figure 9.9).

Figure 9.9 The number of women (top panel) and men (bottom panel) who attended at least once in the analysis period for non-visible haematuria where the event was recorded using Read codes (dark yellow) and text-only (light yellow) in bladder cancer cases (women: *n* = 86/1,352; men: *n* = 226/3,563) and controls (women: *n* = 16/6,266; men: *n* = 44/15,452) separately



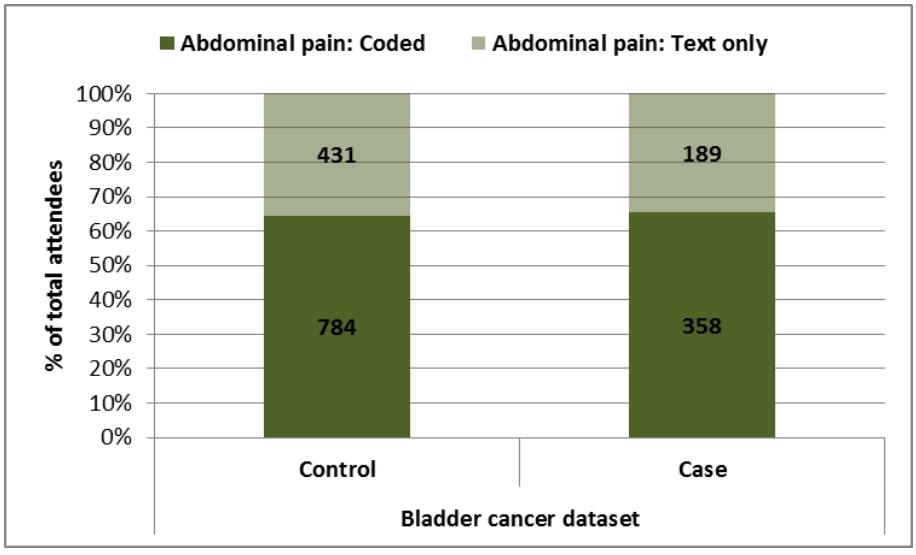


The association led to greater loss of records in hidden text for the female *cases* compared with controls. Of the 16 female controls with at least one episode of non-visible haematuria, the records for 7 (43.8%) were made solely in the hidden text. This proportion *increased* in the female cases, where text-only recording was used for 60/86 (69.8%) women with at least one episode of non-visible haematuria.

9.2.2.1.3 Abdominal pain

There was no evidence of an association between patient status (case or control) and recording style used to identify attendees for abdominal pain ($\chi 2$ test, all patients combined: $p = 0.7$; male patients only: $p = 0.98$; female patients only: $p = 0.5$) (Figure 9.10).

Figure 9.10 The number of patients who attended at least once in the analysis period for abdominal pain where the event was recorded using Read codes (dark green) or text-only (light green) in bladder cancer cases (*n* = 547/4,915) and controls (*n* = 1,215/21,718) separately



Over all patients, coded abdominal pain records predominated, identifying similar proportions of the cases (358/547 = 65%) and controls (784/1,215 = 65%) who had attended for abdominal pain during the analysis period.

9.2.2.2 In the pancreatic cancer dataset

The recording styles that identified attendees for jaundice and abdominal pain in the pancreatic cancer dataset are plotted in Figure 9.11 and Figure 9.12, respectively, The raw data are reported in Table 9.7.

Table 9.7 Numbers of patients attending at least once in the analysis period with abdominal pain or jaundice grouped by recording style

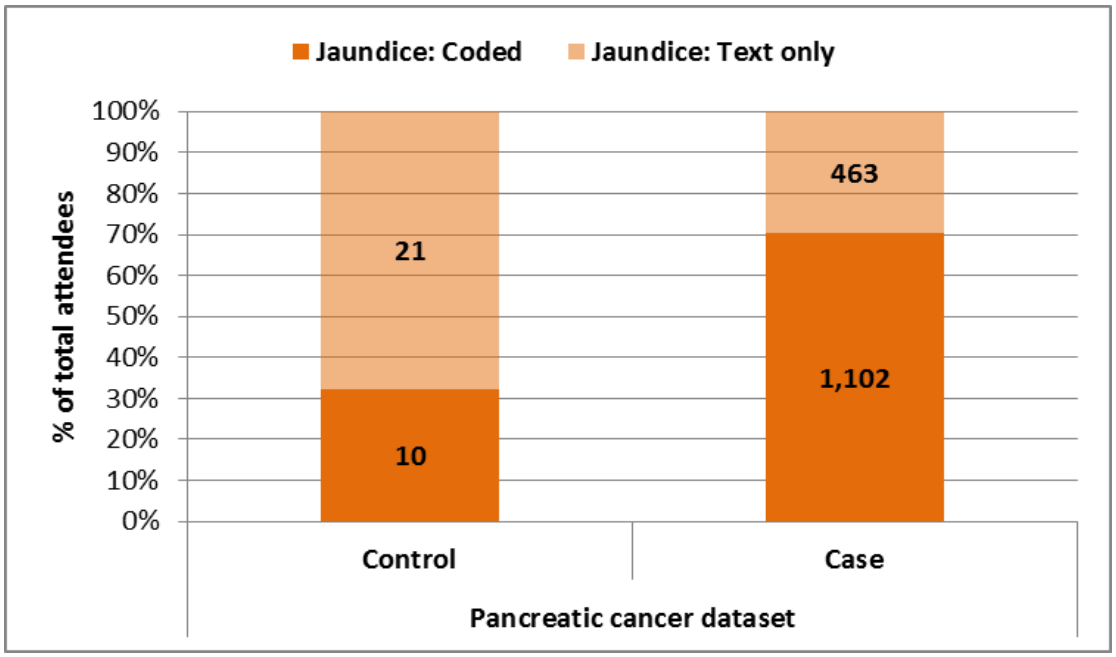
Clinical feature and its recording style	Number of cases (*n* = 3,635) and controls (*n* = 16,459) attending ≥ 1 in the analysis period for the feature		Total, n (% of total with the symptom)
	Cases, *n* (%)	Controls, *n* (%)	
Jaundice	1,102 (30.3)	10 (0.06)	1,112 (69.7)
Jaundice	463 (12.7)	21 (0.1)	484 (30.3)
Total: Jaundice	1,565 (43.1)	31 (0.2)	1,596 (100)
Abdominal pain	1,527 (42.0)	996 (6.1)	2,523 (78.3)
Abdominal pain	383 (10.5)	316 (1.9)	699 (21.7)
Total: Abdominal pain	1,910 (52.5)	1,312 (8.0)	3,222 (100)

9.2.2.2.1 Jaundice

There was strong evidence of an association between patient status (case or control) and the recording style that identified attendees for jaundice ($\chi 2$ test,

p<0.0001) (Figure 9.11). Coded `jaundice` records identified a greater proportion of the cases (1,102/ 1,565 cases = 70%) than of the controls (10/31 = 32%) who had attended for jaundice in the analysis period.

Figure 9.11 The number of patients who attended at least once in the analysis period for jaundice where the event was recorded using Read codes (dark orange) or text-only (light orange) in pancreatic cancer cases (*n* = 1,565/3,635) and controls (*n* = 31/21,718) separately

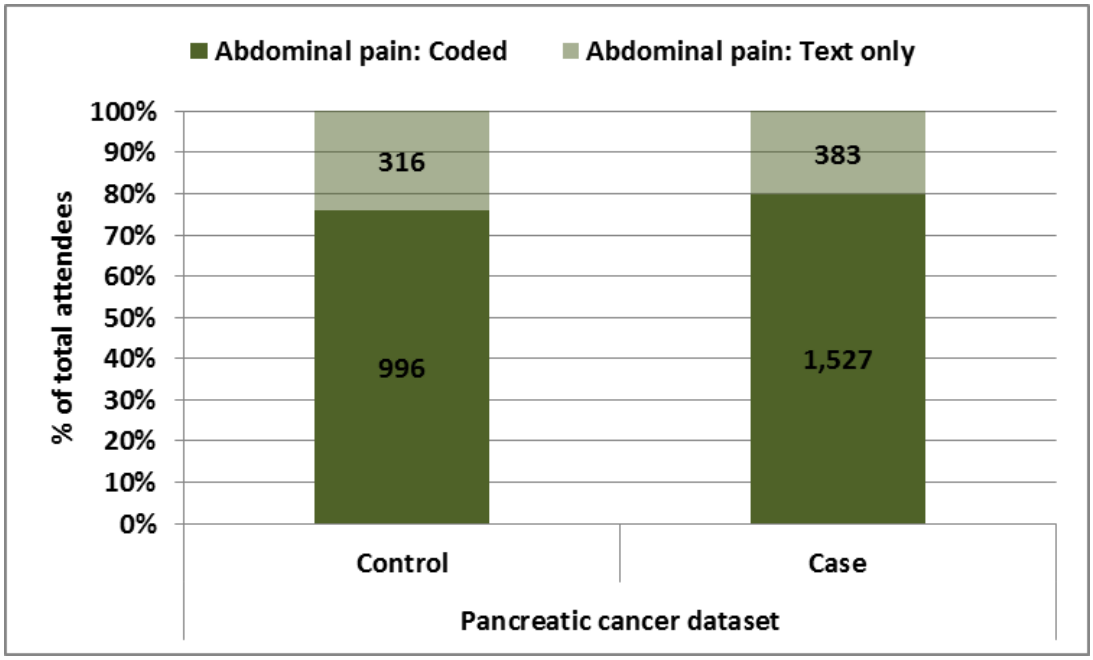


The association between recording style of jaundice and patient status was similar in men as in women. In men, coding identified 533/769 (69.3%) of the cases and 6/16 (37.5%) of the controls (χ2 test, p<0.0001). Similarly, in women, coding identified 569/796 (71.5%) of the cases and 4/15 (26.7%) of the controls (χ2 test, p<0.0001).

9.2.2.2.2 Abdominal pain

There was also strong evidence of an association between patient status and recording style used to identify attendees for abdominal pain (χ2 test, p<0.001). Coded `abdominal pain` records predominated overall, but identified a slightly greater proportion of the cases (1,527/1,910 = 80%) than of the controls (996/1,312 = 76%) who had attended for abdominal pain during the analysis period (see Figure 9.12). Despite its statistical significance, the size of the preference was very small and of questionable clinical relevance.

Figure 9.12 The number of patients who attended at least once in the analysis period for abdominal pain where the event was recorded using Read codes (dark green) or text-only (light green) in pancreatic cancer cases (*n* = 1,910/3,635) and controls (*n* = 1,312/16,459) separately



In men, there was no association between recording style of abdominal pain and patient status (χ2 test, p = 0.3). Coding identified 750/927 (80.9%) of the

male cases and 431/548 (78.7%) of the male controls. In contrast, there was evidence of an association between recording style of abdominal pain and patient status in the women ($\chi 2$ test, $p<0.05$). Coding identified a slightly greater proportion of female cases with at least one episode of abdominal pain (777/983 = 79.0%), compared with 565/764 (74.0%) of the female controls. Again, despite its statistical significance, the size of the preference was very small and of questionable clinical relevance.

9.2.3 Association between recording style and clinical context of symptom presentation

In this analysis, the aim was to compare recording styles between cancer datasets to address research question 4: ‘Does the recording style vary with the clinical context of presentation of a symptom?’ (see Section 5).

Three symptom thesauri of Read codes – one each for haematuria, jaundice and abdominal pain – were used to compare recording styles (coded or text-only) between bladder and pancreatic cancer datasets (see Section 6.8 for methods). For haematuria and jaundice, respectively, I used the symptom thesauri from the bladder and pancreatic cancer studies. For abdominal pain, I used a ‘generic’ thesaurus consisting of a comprehensive list of all forms of abdominal pain rather than either of the bespoke thesauri created in the original studies (see Appendix 4: Symptom thesauri). For this reason, in this section the numbers of patients identified as attending for abdominal pain vary from those reported in the original studies.

9.2.3.1 General recording style

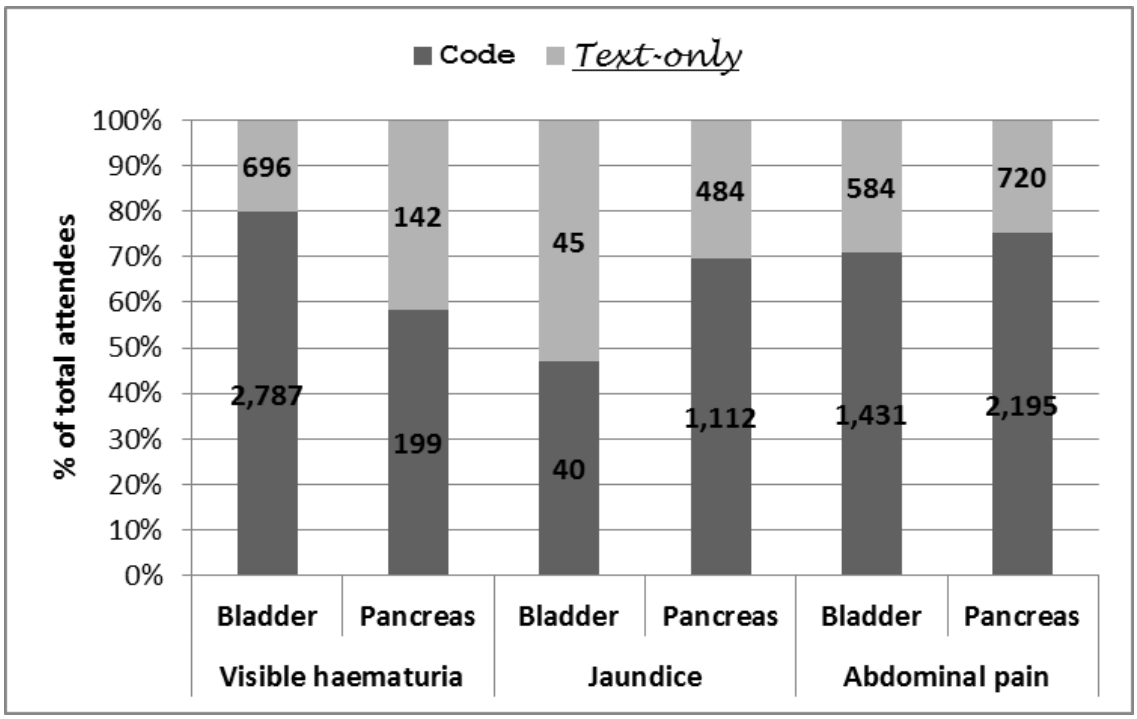
The numbers of patients who attended at least once in the analysis period for each of the features is reported in Table 9.8 and in Figure 9.13, grouped by recording style.[a]

Table 9.8 The effect of the context of presentation on symptom recording style for alarm and non-alarm symptoms. The same symptom thesauri were used to obtain data from both datasets

Symptom	Cancer site	Number (% of total) of patients attending at least once where the events were recorded using:		
		Read code, *n* (%)	*Text-only*, *n* (%)	**Total**, *n* (%)
Visible haematuria	Bladder	2,787 (80)	696 (20)	3,483 (100)
	Pancreas	199 (58)	142 (42)	341 (100)
Jaundice	Bladder	40 (47)	45 (53)	85 (100)
	Pancreas	1,112 (70)	484 (30)	1,596 (100)
Abdominal pain	Bladder	1,431 (71)	584 (29)	2,015 (100)
	Pancreas	2,195 (75)	720 (25)	2,915 (100)

[a] Remember, that text-only is reserved for patients whose attendances were never recorded using a code.

Figure 9.13 Symptom recording styles for alarm and non-alarm symptoms in the bladder and pancreatic cancer datasets. The same symptom thesauri were used to obtain data from both datasets



9.2.3.1.1 Visible haematuria

Coded `visible haematuria` records identified a greater proportion of attendees compared with text-only *visible haematuria* records in both cancer datasets (Table 9.8 and Figure 9.13). While this preference was very strong in the bladder cancer dataset (2,787/3,483 = 80%), it was marginal in the pancreatic cancer dataset (199/341 = 58%).

9.2.3.1.2 Jaundice

In the pancreatic cancer dataset, coded `jaundice` records (1,112/1,596 = 70%) identified more than twice the number of attendees for jaundice compared

with text-only *jaundice* records (484/1,596 = 30%). In the bladder cancer dataset, in contrast, coded `jaundice` (40/85 = 47%) and text-only *jaundice* records (45/85 = 53%) identified similar proportions of attendees.

9.2.3.1.3 Abdominal pain

Coded `abdominal pain` records identified the majority of attendees with this symptom in both datasets; namely 1,431 out of 2,015 attendees (71%) in the bladder, and 2,195/2,915 (75%) attendees in the pancreatic, cancer dataset.

9.2.3.2 Association between recording style and case/control status for rare and common symptoms

The results in this section relate to research questions 2: ‘Are studies of coded data vulnerable to bias arising from the differential use of text and codes between comparison groups?’ and 4: ‘Does the recording style vary with the clinical context of presentation of a symptom?’ (see Section 5).

The association between patient status and recording style reported in Section 9.2.2 did not hold when the context of symptom presentation was changed. The propensity for coded records to identify a greater proportion of cases than controls attending for alarm symptoms only occurred when the symptoms were presented in the context of their associated cancer.

9.2.3.2.1 Visible haematuria

To recap, previously reported results (Section 9.2.3.1.1) indicated that, at the overall patient level, coded `visible haematuria` records identify a greater proportion of the attendees for this symptom compared with text-only *visible haematuria* records. This effect was more marked in the bladder (2,787/3,483 = 80%) than in the pancreatic (199/341 = 58%) cancer dataset (see Figure 9.13 and Table 9.8).

My analysis in the bladder cancer dataset also showed that, at the level of cases and controls, coding identified a greater proportion of cases with visible haematuria compared with controls (see Figure 9.5). In contrast, in the pancreatic cancer dataset, there was marginal evidence of a reversal of the bias, in that coded `visible haematuria` records identified a greater proportion of *controls* than cases with a history of visible haematuria ($\chi 2$ test, p= 0.05) (Figure 9.14).

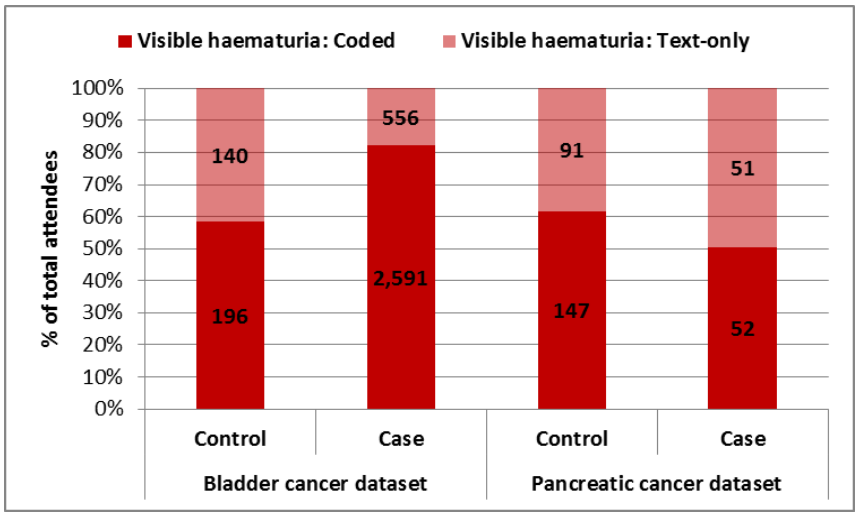
Looking at the pancreatic cancer controls first, the pattern was similar to that observed in the bladder cancer dataset. Coded `visible haematuria` records identified 147 of 238 controls (63%) with a history of the symptom, whereas text-only *visible haematuria* records identified only 91 of 238 (38%) attendees (see Table 9.9 and Figure 9.14). Therefore, this reversal in bias was driven wholly by a change in recording pattern for cases. Indeed, coded `visible haematuria` records identified a similar proportion of the cases (51/103=50%) as did the text-only *visible haematuria* records (52/103=50%).

Table 9.9 Numbers of patients presenting at least once with visible haematuria in the analysis period in the bladder and pancreatic datasets. The recording style (Read code or text only) is reported for cases and controls separately. The same symptom thesaurus was used for both cancer sites

Cancer site	Patient status	Number (%) of total) of patients attending at least once estimated using:		
		Visible haematuria	*Visible haematuria*	Visible haematuria
Bladder***	Control	196 (58)	140 (42)	336 (100)
	Case	2,591 (82)	556 (18)	3,147 (100)
	Total	*2,787 (80)*	*696 (20)*	*3,483 (100)*
Pancreas*	Control	147 (62)	91 (38)	238 (100)
	Case	52 (50)	51 (50)	103 (100)
	Total	*199 (58)*	*142 (42)*	*341 (100)*

***p< 0.0001, *p= 0.05; χ2 test of the null hypothesis that there is no association between patient status and recording style.

Figure 9.14 Visible haematuria recording style in cases and controls compared in the contexts of bladder (left) and pancreatic (right) cancers



9.2.3.2.2 Jaundice

As reported in Section 9.2.3.1.2, coded `jaundice` records (1,112/1,596 = 70%) identified a greater proportion of attendees for this symptom compared with text-only *jaundice* records (484/1,596 = 30%). This pattern was not apparent in the bladder cancer dataset (`jaundice`: 40/85 = 47%, *jaundice*: 45/85 = 53%). The propensity for coded records to identify attendees was more marked in patients who later transpired to have pancreatic cancer compared with controls (1,102/1,565 = 70% in cases vs 10/31 = 32% in controls; $\chi 2$ test, $p<0.0001$) (see Section 9.2.2.2.1, see also Table 9.10 and Figure 9.15).

In contrast, in the bladder cancer dataset, there was marginal evidence suggesting a reversal of this bias. Coded `jaundice` records identified a greater proportion of *controls* than cases with a history of the symptom, although it should be noted that the numbers are small ($\chi 2$ test, $p = 0.05$).

The reversal in bias was driven by both a loss of overall tendency for either recording style in the controls (`jaundice`: 30/64 = 47% vs *jaundice*: 34/64 = 53%) and a loss of the tendency for coded `jaundice` records to identify cases (6/21 = 29%).

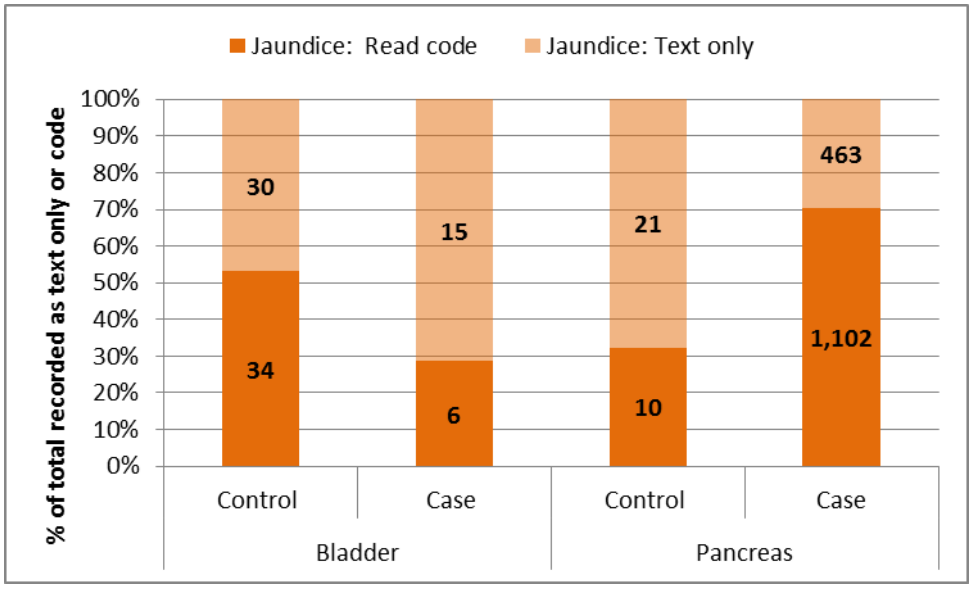
The recording style patterns were not the same in the control groups for the bladder and pancreatic cancer datasets, which may be accounted for by the small numbers of these patients attending for jaundice.

Table 9.10 Numbers of patients presenting at least once with jaundice in the analysis period in the bladder and pancreatic datasets. The recording style (Read code or text only) is reported for cases and controls separately. The same symptom thesauri were used for both cancer sites

Cancer site	Patient status	Number of patients attending at least once estimated using:		
		Jaundice	*Jaundice*	Jaundice
Bladder*	Control	34 (53)	30 (47)	64 (100)
	Case	6 (29)	15 (71)	21 (100)
	Total	*40 (47)*	*45 (53)*	*85 (100)*
Pancreas***	Control	10 (32)	21 (68)	31 (100)
	Case	1,102 (70)	463 (30)	1,565 (100)
	Total	*1,112 (70)*	*484 (30)*	*1,596 (100)*

***p< 0.0001, *p= 0.05; $\chi 2$ test of the null hypothesis that there is no association between patient status and recording style.

Figure 9.15 Jaundice recording style in cases and controls compared in the contexts of bladder (left) and pancreatic (right) cancers



9.2.3.2.3 Abdominal pain

As reported in Section 9.2.3.1.3, coded `abdominal pain` records identified a greater proportion of attendees for this symptom compared with text-only *abdominal pain* records, an effect that was similar in the bladder (1,431/2,015 = 71%) and pancreatic (2,195/2,915 = 75%) cancer datasets.

As described above (see Section 9.2.3.1.3), this effect was slightly more marked in patients who later transpired to have pancreatic cancer (1,527/1,910 = 80%) compared with controls (996/1,312 = 76%) ($\chi 2$ test, $p<0.001$) (see Table 9.11 and Figure 9.16).

In contrast, there was no association between patient status and the recording style identifying attendees for abdominal pain in the bladder cancer dataset ($\chi 2$ test, $p=0.4$). Coded `abdominal pain` records identified similar proportions of cases (419/600 = 70%) and controls (1,012/1,415 = 72%) (see Table 9.11 and Figure 9.16).

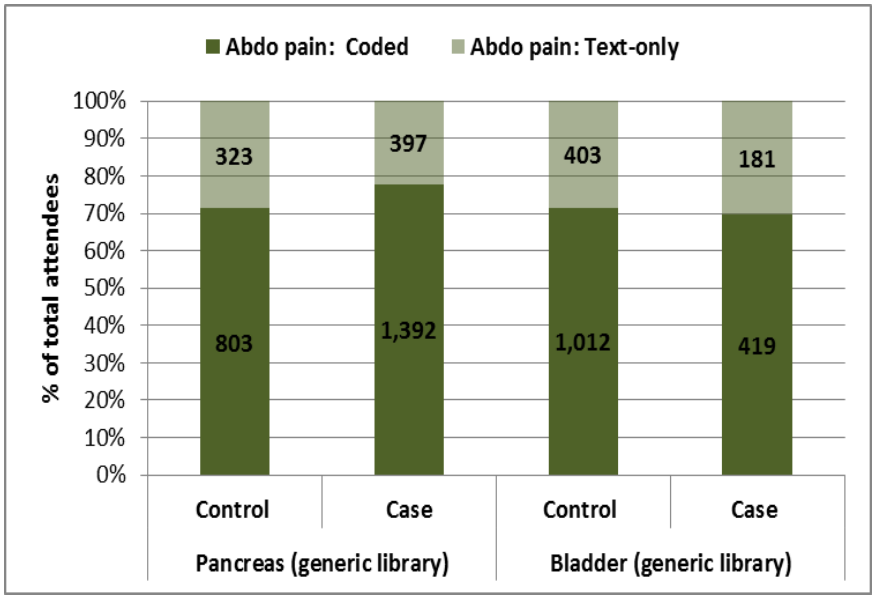
The recording style pattern for abdominal pain was similar in the pancreatic and bladder cancer control groups – mirroring the results for visible haematuria recording style.

Table 9.11 Numbers of patients presenting at least once with abdominal pain in the analysis period in the bladder and pancreatic datasets. The recording style (code or text only) is reported for cases and controls separately. The same symptom thesaurus was used for both cancer sites

Cancer site	Patient status	Number of patients attending at least once estimated using:		
		Abdominal pain	*Abdominal pain*	Abdominal pain
Bladder†	Control	1,012 (72)	403 (28)	1,415 (100)
	Case	419 (70)	181 (30)	600 (100)
	Total	*1,431 (71)*	*584 (29)*	*2,015 (100)*
Pancreas***	Control	803 (71)	323 (29)	1,126 (100)
	Case	1,392 (78)	397 (22)	1,789 (100)
	Total	*2,195 (75)*	*720 (25)*	*2,915 (100)*

***p< 0.001, †p= 0.4; χ2 test of the null hypothesis: there is no association between patient status and recording style.

Figure 9.16 Abdominal pain recording style in cases and controls compared in the contexts of pancreatic and bladder cancers. The same symptom thesaurus was used in both datasets



10 Results: effect of recording style bias on risk estimates for cancer

10.1 Bladder cancer

10.1.1 Estimates of positive predictive value and positive likelihood ratio from coded records

The features of bladder cancer presentation in primary care were identified in the original study extended by my PhD. The numbers of cases and controls with coded records of attendance for each of these features is reported in Table 10.1. Also presented are the positive likelihood ratio and positive predictive value (PPV) for bladder cancer.[a] The results are from my re-construction and re-analysis, and numbers of patients and positive likelihood ratios largely matched[b] those of the original study.[4] The positive predictive values for the

[a] As defined before the International Classification of Disease (ICD) for Oncology created separate diagnostic codes for carcinoma *in situ* of bladder (D09.0) and neoplasm of uncertain behaviour of bladder (D41.4), removing them from code C67 (bladder cancer).

[b] The discrepancies in numbers of patients were essentially trivial, and reflected very slight differences in definitions of some variables, such as that for raised inflammatory markers. Furthermore, I assumed that two similar codes on the same day (e.g. Read codes K197300 'Frank haematuria' and K197199 'Painful haematuria') reflected a single event. The positive predictive values reported in the original study are lower than those reported in my PhD because they were estimated using a prior odds for the post-1998 definition of bladder cancer.

post-1998 definition of bladder cancer are given in an appendix (see Appendix 6: Risk estimates for the post-1998 definition of bladder cancer (C67)).

In terms of frequency of attendance, `visible haematuria` was the most notable feature, with coded records indicating that 2,591/4,915 (52.7%) of cases attended at least once in the analysis period. `Urinary tract infection` was the next most significant feature (in 835/4,915 = 17.0% of cases), followed by `raised creatinine` (in 660/4,915 = 13.4% of cases) and then `dysuria` (in 444/4,915 = 9.0% of cases). `Non-visible haematuria` was recorded as a code in only 127 (2.6%) of cases and 26 (0.1%) of controls (data not shown). This feature therefore failed to meet the 5% threshold for inclusion in univariable analysis (the justification for setting this threshold at 5% is given in Section 6.12.1).

Of all the features, `visible haematuria` was the most strongly predictive of bladder cancer, associated with a 3.98% (3.47–4.571%) chance of the disease. Indeed, patients with bladder cancer were 58.41 (50.69–67.32) times more likely to have visible haematuria compared with controls. However, in all age groups combined, the positive likelihood ratio values and PPVs for all other features, when considered in isolation, were unremarkable.

Table 10.1 The positive likelihood ratio and PPV for bladder cancer (pre-1998 definition) in patients aged ≥40 years presenting with clinical features associated with the disease (codes)

Feature of bladder cancer	Number of patients attending at least once:		Positive likelihood ratio (95% CI)	PPV (%) (95% CI)[ab]
	Cases n (% of *n* = 4,915 cases)	Controls, n (% of *n* = 21,718 controls)		
Abdominal pain	358 (7.3)	784 (3.6)	2.01 (1.79–2.28)	0.14 (0.13–0.16)
Constipation	286 (5.8)	708 (3.3)	1.78 (1.56–2.04)	0.13 (0.11–0.14)
Visible haematuria	2,591 (52.7)	196 (0.9)	58.41[c] (50.69–67.32)	3.98 (3.47–4.57)
Dysuria	444 (9.0)	209 (1.0)	9.39 (7.99–11.03)	0.66 (0.56–0.78)
Urinary tract infection	835 (17.0)	705 (3.2)	5.23 (4.76–5.76)	0.37 (0.34–0.41)
Raised inflammatory markers	332 (6.7)	809 (3.7)	1.81 (1.58–2.06)	0.13 (0.11–0.15)
High white cell count	251 (5.1)	401 (1.8)	2.75 (2.36–3.22)	0.19 (0.17–0.23)
Raised creatinine	660 (13.4)	1,668 (7.7)	1.75 (1.61–1.90)	0.12 (0.11–0.13)

[a] PPV values are adjusted for the consulting population (see Section 6.12.3.5).
[b] PPV values estimated using Bayes' Theorem (see Section 6.12.3.5), assuming a prior odds of 0.000646 based on 2008 UK national incidence data. See Section 15.5.1 for a worked example of how to calculate PPV using Bayes' theorem.
[c] Positive likelihood ratio >10.

10.1.2 Effect of supplementation with text-only records on positive likelihood ratio and PPV

Detailed analysis of GPs' choice of recording style was presented in Section 9 (see Table 9.6). This section focuses on whether supplementing codes with text records alters the risk estimates for bladder cancer in patients with haematuria or abdominal pain. This relates to research questions 2 (Are studies of coded data vulnerable to bias arising from the differential use of text and codes between comparison groups?) and 5 (Do the text data provide additional value to coded data?) (see Section 5).

The numbers of patients attending at least once for abdominal pain or haematuria obtained from coded records in the original study were updated with information from the text. These revised binary `abdominal pain`, `visible haematuria` and `non-visible haematuria` variables were used to re-examine the predictive power of these symptoms for bladder cancer (Table 10.2). Where earlier results suggested that the frequency of text-only records was different in men than in women, revised risk estimates were calculated for men and women separately.

Table 10.2 Positive likelihood ratio and PPV for bladder cancer in patients of all ages presenting with abdominal pain, visible or non-visible haematuria estimated from coded and from coded plus text-only records

Clinical feature and recording style	Number of patients attending at least once: Cases *n* (% of *n* = 4,915 cases)	Number of patients attending at least once: Controls, *n* (% of *n* = 21,718 controls)	Positive likelihood ratio (95% CI)	PPV (%) (95% CI) [a]
Abdominal pain	358 (7.3)	784 (3.6)	2.02 (1.79–2.28)	0.14 (0.13–0.16)
Abdominal pain	542 (11.0)	1,217 (5.6)	1.97 (1.79–2.17)	0.14 (0.13–0.15)
Visible haematuria	2,591 (52.7)	196 (0.9%)	58.41 (50.69–67.32)	3.98 (3.47–4.57)
Visible haematuria	3,147 (64.0)	336 (1.5)	41.39 (37.14–46.11)	2.85 (2.57–3.17)
Non-visible haematuria	127 (2.6)	26 (0.1)	n/a	n/a
Non-visible haematuria	312 (6.3)	60 (0.3)	22.98 (17.46–30.24)	1.60 (1.22–2.10)

[a] PPV values are adjusted for the consulting population (see Section 6.12.3.5). PPV values estimated using Bayes' Theorem (see Section 6.12.3.5), assuming a prior odds of 0.000646 based on 2008 UK national incidence data.

10.1.2.1 Abdominal pain

The 95% confidence intervals suggested that supplementing coded abdominal pain records with text-only records did not alter the positive likelihood ratio (abdominal pain: 2.02, 95% CI: 1.79–2.28 vs abdominal pain: 1.97, 95% CI: 1.79–2.17) or the PPV (abdominal pain: 0.14, 95% CI: 0.13–0.15 vs abdominal pain: 0.14, 95% CI: 0.13–0.15) (Table 10.2). Risk estimates were similar in men and women, and were unaffected by addition of text-only records (data not reported).

10.1.2.2 Visible haematuria

In contrast, supplementing the coded visible haematuria record with text-only records reduced the positive likelihood ratio from 58.41 (95% CI: 50.69–67.32) (visible haematuria) to 41.39 (95% CI: 37.14–46.11) (visible haematuria). Similarly, the PPV fell from 3.98% (95% CI: 3.47–4.57%) (visible haematuria) to 2.85% (95% CI: 2.57–3.17%) (visible haematuria) (Table 10.2).

The risk of bladder cancer associated with visible haematuria is reported for men and women separately in Table 10.3. The estimates of PPV in men and in women from coded records were greater than those estimated for men and women combined. For men, this was due to their greater prior odds for bladder cancer (0.001002 for men compared with 0.000646 for men and women combined). For women, this was due to their relatively high likelihood ratio

(155.14, 95% CI: 98.67–243.92 in women compared with 58.41, 95% CI: 50.69–67.32 for all patients combined).

Table 10.3 Positive likelihood ratio and PPV for bladder cancer in men and women separately, aged over 40 years, presenting with visible haematuria estimated from coded and from coded plus text-only records

Visible haematuria recording style	Number (%) of patients attending at least once:		Positive likelihood ratio (95% CI)	PPV (%) (95% CI) [a]
	Cases, *n* (% of cases)	Controls, *n* (% of controls)		
Men (*n*=3,563 cases, *n*=15,452 controls)				
Visible haematuria	1,955 (54.9)	177 (1.1)	47.90 (41.25–55.62)	5.14 (4.46–5.93)
Visible haematuria	2,356 (66.1)	279 (1.8)	36.62 (32.53–41.23)	3.97 (3.54–4.46)
Women (*n*=1,352 cases, *n*=6,266 controls)				
Visible haematuria	636 (47.0)	19 (0.3)	155.14 (98.67–243.92)	5.10 (3.30–7.80)
Visible haematuria	791 (58.5)	57 (0.9)	64.32 (49.48–83.60)	2.17 (1.68–2.81)

[a] PPV values are adjusted for the consulting population (see Section 6.12.3.5). PPV values estimated using Bayes' Theorem (see Section 6.12.3.5), assuming prior odds of 0.001002 for men and of 0.000323 for women, based on 2008 UK national incidence data.

Looking at the 95% confidence intervals, addition of text-only records of attendance for visible haematuria reduced the risk estimates both in men and women. However, the extent of the correction brought about by addition of text-only records was much more marked in the women than in the men. The PPV from coded records in men was 5.14% (95% CI: 4.46–5.93%), and addition of text-only records reduced the PPV to 3.97% (95% CI: 3.54–4.46%). In women, however, the PPV from coded records was 5.10% (95% CI: 3.30–7.80%), and was reduced to 2.17% (95% CI: 1.68–2.81%) after addition of text-only records.

10.1.2.3 Non-visible haematuria

The probability of bladder cancer in patients according to `non-visible haematuria` was 1.60% (95% CI: 1.22–2.10%), and cases were approximately 23 times more likely to experience non-visible haematuria than controls (likelihood ratio 22.98, 95% CI: 17.46–30.24) (Table 10.2). There were no comparison data from the original study, as non-visible haematuria did not meet the 5% threshold for inclusion in analysis.

10.1.3 Estimates of odds ratios in univariable analyses from coded records

The frequency of attendance for each feature included in the final model of the original study[4] is reported in Table 10.4 along with its odds ratio for disease in univariable analysis. The results are from my re-construction and re-analysis, and largely matched those of the original study.

Table 10.4 Frequency of attendance for features included in the final model of the original study,[4] plus the univariable analysis

Feature of bladder cancer	*n*[a]	Number of participants		Odds ratio in univariable analysis (95% CI)	p[b]
		Cases, *n* (% of *n* = 4,915 cases)	Controls, *n* (% of *n* = 21,718 controls)		
Abdominal pain	0	4,557 (92.7)	20,934 (96.4)	–	–
	≥1	358 (7.3)	784 (3.6)	2.1 (1.9–2.4)	0.0001
Constipation	0	4,629 (94.2)	21,010 (96.7)	–	–
	≥1	286 (5.8)	708 (3.3)	1.9 (1.6–2.2)	0.0001
Visible haematuria[c]	0	2,324 (47.3)	21,522 (99.1)	–	–
	1	1,301 (26.5)	133 (0.6)	100.5 (78.1 – 129.3)	0.0001
	2	721 (14.7)	39 (0.2)	240.4 (157.8–366.5)	0.0001
	≥3	568 (11.6)	24 (0.11)	308.6 (186.4–510.9)	0.0001
Dysuria	0	4,471 (91.0)	21,509 (99.0)	–	–
	≥1	444 (9.0)	209 (1.0)	10.6 (8.9–12.6)	0.0001
Urinary tract infection[d]	0	4,080 (83.0)	21,013 (96.8)	–	–
	1	511 (10.4)	503 (2.3)	5.9 (5.1–6.7)	0.0001
	≥2	324 (6.6)	202 (0.9)	10.1 (8.3–12.3)	0.0001
Raised inflammatory markers	0	4,583 (93.3)	20,909 (96.3)	–	–
	≥1	332 (6.7)	809 (3.7)	2.0 (1.8–2.3)	0.0001
High white cell count	0	4,664 (94.9)	21,317 (98.2)	–	–
	≥1	251 (5.1)	401 (1.8)	3.1 (2.6–3.6)	0.0001
Raised creatinine	0	4,255 (86.6)	20,050 (92.3)	–	–
	≥1	660 (13.4)	1,668 (7.7)	2.1 (1.9–2.3)	0.0001

Notes: 95% CI = 95% confidence interval.
[a] *n* is the number of times the patient attended for the feature in the analysis period.
[b] Wald test of the null hypothesis that the odds ratio is 1.

[c] Strong evidence that the association between visible haematuria and bladder cancer varied with number of attendances (χ^2= 1,794.5, 3 d.f., p<0.0001).
[d] Strong evidence that the association between urinary tract infection and bladder cancer varied with number of attendances (χ2= 22.6, 1 d.f., p<0.0001).

10.1.3.1 Strength of association with bladder cancer

Visible haematuria was the most significant feature in terms of strength of association with bladder cancer in univariable analysis. The odds of bladder cancer in patients who attended only once with visible haematuria was 100.5 times (95% CI: 78.1–129.3, Wald test p<0.0001) higher than in patients for whom there was no evidence from codes that they had attended for this symptom. Dysuria had the second strongest association with bladder cancer (odds ratio = 10.6, 95% CI: 8.9–12.6 for at least one presentation, p<0.0001), followed by urinary tract infection (odds ratio = 5.9, 95% CI: 5.1–6.7 for a single attendance, p<0.0001). Abdominal pain, constipation, a high white cell count, a high creatinine level or raised inflammatory markers were all less strongly associated with bladder cancer, with odds ratios at or below 3 for at least one attendance.

10.1.3.2 The significance of re-attendance for a feature

There was strong evidence that re-attendance for visible haematuria was significantly associated with cancer. The association between bladder cancer and visible haematuria varied with the number of attendances in the analysis period (χ^2=1,794.5, 3 d.f., Wald test p<0.0001), with further episodes increasing the odds of disease. Similarly, the odds of bladder cancer increased with repeat attendance for urinary tract infection (χ^2=22.6, 1 d.f.,

p<0.0001). The significance of re-attendance for abdominal pain, constipation or dysuria could not be assessed, as too few patients attended more than once in the analysis period. Similarly, GPs did not re-order tests in enough patients for inflammatory markers, creatinine or a white cell count to allow assessment of the significance of repeat testing.

10.1.4 Effect on odds ratios in univariable analyses of supplementation with text-only records

The odds ratios for bladder cancer after addition of text-only records of attendance for non-visible and visible haematuria and for abdominal pain are reported below.

10.1.4.1 Non-visible haematuria

Descriptors in the text allowed separate identification of visible and non-visible forms of haematuria, enabling re-assessment of the frequency of attendance for the latter (see Table 10.5). As described above, there were insufficient coded non-visible haematuria records to meet the threshold for inclusion in univariable analysis. Similarly, text-only *non-visible haematuria* records failed the 5% threshold for inclusion in univariable analysis, amounting to 184 of 4,915 (3.8%) of cases. Despite this, non-visible haematuria and *non-visible haematuria were both* significantly associated with bladder cancer (Table 10.5), with respective odds ratios of 28.4 (95% CI 17.5–46.1, Wald test,

p<0.0001) and 27.7 (95% CI: 18.8–40.8) (p<0.0001) for at least one attendance.

Table 10.5 Numbers of cases and controls attending with non-visible haematuria in the analysis period, according to the recording style, and independent association with bladder cancer in univariable analysis

Recording style [a]	*n* [b]	Number (%) of patients presenting with non-visible haematuria		Odds ratio (95% CI) in univariable analysis	p value [c]
		Cases (*n* = 4,915)	Controls (*n* = 21,718)		
Non-visible haematuria	0	4,788 (97.4)	21,692 (99.9)	–	–
	≥1	127 (2.6)	26 (0.1)	28.4 (17.5–46.1)[d]	0.0001
Non-visible haematuria	0	4,730 (96.2)	21,684 (99.8)	-	-
	≥1	185 (3.8)	34 (0.2)	27.7 (18.8–40.8)[d]	0.0001
Non-visible haematuria	0	4,603 (93.6)	21,658 (99.7)	–	–
	≥1	312 (6.4)	60 (0.3)	28.0 (20.7–37.9)	0.0001

Notes:
[a] Non-visible haematuria – some or all attendances recorded as a code; *Non-visible haematuria* – all attendances recorded as text only; Non-visible haematuria – total, i.e. coded record supplemented with text-only records.
[b] Number of attendances in the analysis period.
[c] Wald test that the odds ratio is 1.
[d] No evidence that association between non-visible haematuria and bladder cancer varies with recording style ($\chi 2 = 0.01$, 1 d.f., $p = 0.936$).

There was no evidence that the association between bladder cancer and non-visible haematuria varied with recording style (Wald test, $\chi^2 = 0.01$, 1 d.f.. p = 0.936).

Supplementation with text-only records increased the total number of patients identified as attending at least once for non-visible haematuria to 312 (6.4% of 4,915 cases). Thus, the 5% threshold was exceeded, permitting estimation of the odds ratio in univariable analysis (OR 28.0, 95% CI: 20.7–37.9, p<0.0001). Therefore, non-visible haematuria was later included as an explanatory variable in multivariable analysis.

10.1.4.2 Visible haematuria

There was also significant recording of *visible haematuria* in text-only records (Table 10.6), easily exceeding the 5% threshold for inclusion in univariable analysis. Of the 4,915 cases, 556 (11.3%) attended at least once where this was never documented as a code. The text-only records indicated that patients did re-attend for visible haematuria; however, the numbers were too low to permit meaningful analysis of its significance in terms of association with bladder cancer.

In contrast with non-visible haematuria, there was strong evidence that the association between visible haematuria and bladder cancer varied with the recording style ($\chi2 = 75.8$, 3 d.f., p<0.0001). The odds ratio for bladder cancer from *visible haematuria* records was 21.0 (95% CI: 17.1–25.6) (Wald test, p<0.0001) for at least one attendance. However, it was 100.5 (95% CI: 78.1–

129.3) for a single attendance according to the coded visible haematuria records.

Table 10.6 Numbers of cases and controls attending with visible haematuria in the analysis period, according to the recording style, and independent association with bladder cancer in univariable analysis

Recording style [a]	n_a [b]	Number (%) of patients attending:		Odds ratio (95% CI)	p value [c]
		Cases (*n* = 4,915)	Controls (*n* = 21,718)		
Visible haematuria	0	2,324 (47.3)	21,522 (99.1)	–	–
	1	1,301 (26.5)	133 (0.6)	100.5 (78.1–129.3)	0.0001
	2	721 (14.7)	39 (0.2)	240.4 (157.8–366.5)	0.0001
	≥3	569 (11.6)	24 (0.11)	308.6 (186.4–510.9)	0.0001
Visible haematuria	0	4,359 (88.7)	21,548 (99.4)	–	–
	≥1	556 (11.3)	140 (0.6)	21.0 (17.1–25.6)	0.0001
Visible haematuria	0	1,768 (36.0)	21,382 (98.5)	–	–
	1	1,306 (26.6)	229 (1.1)	75.9 (60.9–94.6)	0.0001
	2	910 (18.5)	64 (0.3)	223.3 (156.4–318.7)	0.0001
	≥3	931 (18.9)	43 (0.2)	473.6 (306.2–732.3)	0.0001

[a] Strong evidence that the association between bladder cancer and visible haematuria varied with the recording style ($\chi 2$= 75.8, 3 d.f., p<0.0001).
[b] Number of attendances in the analysis period (the maximum number of attendances examined was that level containing no fewer than 5% of cases).
[c] Wald test of the null hypothesis that the OR is 1.

The odds ratio for the combined variable visible haematuria was 75.9 (95% CI: 60.9–94.6, p<0.0001) for a single attendance. The effect of re-attendance with visible haematuria was highly significant ($\chi 2$= 77.97, 2 d.f,

p<0.0001), with the odds ratio rising to 473.6 (95% CI: 306.2–732.3, p<0.0001) for three or more attendances.

10.1.4.3 Abdominal pain

There was a significant amount of text-only *abdominal pain* recording (see Table 10.7), which identified that 189 of the 4,915 (3.8%) cases attended at least once in the analysis period. However, the numbers were insufficient for it to meet the 5% threshold for inclusion in univariable analysis. Despite this, *abdominal pain* *was* independently associated with bladder cancer, with an odds ratio of 2.1 (1.7–2.5) (Wald test, p<0.0001).

There was no evidence that the association between bladder cancer and abdominal pain varied with the recording style ($\chi 2 = 1.0$, 1 d.f., $p = 0.99$). The odds ratios in univariable analysis were similar for `abdominal pain` (2.1, 95% CI: 1.9–2.4, p<0.0001), *abdominal pain* (2.1, 95% CI: 1.7–2.5, p<0.0001) and the combined `abdominal pain` (2.2, 95% CI: 2.0–2.4, p<0.0001) variables. The significance of re-attendance with abdominal pain in the analysis period could not be assessed, owing to small numbers.

Table 10.7 Frequency of attendance for abdominal pain in the analysis period in cases and controls grouped by recording style, and independent association with bladder cancer in univariable analysis

Recording style[a d]	n_a[b]	Number (%) of patients attending for abdominal pain		Odds ratio (95% CI)	p value[c]
		Cases (*n* = 4,915)	Controls (*n* = 21,718)		
Abdominal pain	0	4,557 (92.7)	20,934 (96.4)	–	–
	≥1	358 (7.3)	784 (3.6)	2.1 (1.9–2.4)	0.0001
Abdominal pain	0	4,726 (96.2)	21,287 (98.0)	–	–
	≥1	189 (3.8)	431 (2.0)	2.1 (1.7–2.5)	0.0001
Abdominal pain	0	4,368 (88.9)	20,503 (94.4)	–	–
	≥1	547 (11.1)	1,215 (5.6)	2.2 (2.0–2.4)	0.0001

Notes:
[a] Abdominal pain – some or all attendances recorded as a code; *Abdominal pain* – all attendances recorded as text only; Abdominal pain – total, i.e. coded record supplemented with text-only records.
[b] Number of attendances in the analysis period.
[c] Wald test that the odds ratio is 1.
[d] No evidence that the association between abdominal pain and bladder cancer varies with recording style ($\chi 2 = 1.0$, 1 d.f., $p = 0.99$).

10.1.5 Multivariable analyses

10.1.5.1 The basic model estimated using coded records

The multivariable conditional logistic regression analysis of the presenting features of bladder cancer, as identified in the original study,[4] was re-run using solely code-based variables to replicate the basic model (see Table 10.8).

Visible haematuria was the feature most strongly associated with bladder cancer. According to the coded records, the odds of bladder cancer were 107.5

(95% CI: 82.9–139.3) (Wald test, p<0.0001) times higher in patients who attended once with the symptom compared with those who never attended. The strength of association with bladder cancer varied with further episodes of visible haematuria ($\chi 2 = 25.01$, 2 d.f., p<0.0001), rising to an odds ratio of 345.7 (95% CI: 198.6–601.7) (p<0.0001) for three or more attendances in the analysis period (Table 10.8).

Of the remaining features of bladder cancer, dysuria had the strongest association with bladder cancer, with an odds ratio of 6.7 (95% CI: 5.2–8.7) (Wald test, p<0.0001). Re-attendance for urinary tract infection increased the odds of bladder cancer, from 4.4 (95% CI: 3.6–5.3) to 6.8 (95% CI: 5.1–9.1) ($\chi 2 = 5.7$, 1 d.f., p<0.05) (see Table 10.8).

Abdominal pain was also associated with bladder cancer, albeit less strongly than visible haematuria. The odds ratio for patients attending at least once with abdominal pain was 2.0 (95% CI: 1.7–2.5) (Wald test, p<0.0001) (Table 10.8). Insufficient numbers of patients re-attended with the symptom to investigate the significance of further episodes of abdominal pain.

Table 10.8 Frequency of attendance for features of bladder cancer in the year before analysis and odds ratio in multivariable conditional logistic regression analysis; estimates were obtained from coded records

Feature of bladder cancer		*n*[a]	Number of:		Odds ratio (95% CI)
			Cases, n (% of *n* = 4,915 cases)	Controls, n (% of *n* = 21,718 controls)	
Symptoms	Abdominal pain	0	4,557 (92.7)	20,934 (96.4)	–
		≥1	358 (7.3)	784 (3.6)	2.0 (1.7–2.5)***
	Constipation	0	4,629 (94.2)	21,010 (96.7)	–
		≥1	286 (5.8)	708 (3.3)	1.5 (1.2–1.8)**
	Visible haematuria	0	2,324 (47.3)	21,522 (99.1)	–
		1	1,301 (26.5)	133 (0.6)	107.5 (82.9–139.3)***
		2	721 (14.7)	39 (0.2)	275.0 (176.6–428.2)***
		≥3	569 (11.6)	24 (0.1)	345.7 (198.6–601.7)***
	Dysuria	0	4,471 (91.0)	21,509 (99.0)	–
		≥1	444 (9.0)	209 (1.0)	6.7 (5.2–8.7)***
Disease	Urinary tract infection	0	4,080 (83.0)	21,013 (96.8)	–
		1	511 (10.4)	503 (2.3)	4.4 (3.6–5.3)***
		≥2	324 (6.6)	202 (0.9)	6.8 (5.1–9.1)***
Investigation	Raised inflammatory markers	0	4,583 (93.2)	20,909 (96.3)	–
		≥1	332 (6.8)	809 (3.7)	1.5 (1.2–1.9)***

Feature of bladder cancer		*n*[a]	Number of:		Odds ratio (95% CI)
			Cases, n (% of *n* = 4,915 cases)	Controls, n (% of *n* = 21,718 controls)	
	High white cell count	0	4,664 (94.9)	21,317 (98.2)	–
		≥1	251 (5.1)	401 (1.8)	2.1 (1.6–2.8)****
	Raised creatinine	0	4,255 (86.6)	20,050 (92.3)	–
		≥1	660 (13.4)	1,668 (7.7)	2.0 (1.7–2.3)***

**p<0.001, ***p<0.0001, Wald test of the null hypothesis that the OR is 1.

[a]*n*, Number of attendances in the analysis period.

10.1.5.2 Goodness of fit with and without text-only records

The effect of adding the new, text-based, variables *abdominal pain*, *visible haematuria* and non-visible haematuria to the original model was highly significant (nested models, likelihood ratio $\chi 2$ = 1,420.76, 3 d.f., p<0.0001). Therefore, text-only records were included for all three symptoms in the revised final model (see Table 10.9 below).

10.1.5.3 Effect modification

The original study did not report any significant effect modification in the model; however, this was revisited and sought on clinical grounds. Significant antagonistic effect modification was found between urinary tract infection and each of visible haematuria (Wald test, $\chi 2$ = 128.38, 6 d.f., p<0.0001), dysuria ($\chi 2$ = 14.55, 2 d.f., p<0.001) and non-visible

haematuria ($\chi^2 = 16.97$, 2 d.f., $p<0.001$). The individual odds ratios for effect modification are reported in Table 10.9.

10.1.6 Revised final model

The revised final model, including the text-only records and effect modification, is reported in Table 10.9.

Table 10.9 Conditional logistic regression analysis of the final model of bladder cancer presentation

Feature of bladder cancer		*n*[a]	Number of: Cases (*n* = 4,915)	Controls (*n* = 21,718)	Odds ratio (95% CI)
Symptoms	Abdominal pain	0	4,368 (88.9)	20,503 (94.4)	–
		≥1	547 (11.1)	1,215 (5.6)	2.0 (1.7–2.4)***
	Constipation	0	4,629 (94.2)	21,010 (96.7)	–
		≥1	286 (5.8)	708 (3.3)	1.4 (1.1–1.8)*
	Visible haematuria	0	1,768 (36.0)	21,382 (98.5)	–
		1	1,306 (26.6)	229 (1.1)	96.6 (75.0–124.5) ***
		2	910 (18.5)	64 (0.3)	312.4 (204.6–477.0) ***
		≥3	931 (18.9)	43 (0.2)	644.5 (384.9–1,079.1) ***
	Dysuria	0	4,471 (91.0)	21,509 (99.0)	–
		≥1	444 (9.0)	209 (1.0)	7.9 (5.6–11.1) ***

Disease	Urinary tract infection	0	4,080 (83.0)	21,013 (96.8)	–
		1	511 (10.4)	503 (2.3)	4.7 (3.7–6.1) ***
		≥2	324 (6.6)	202 (0.9)	9.2 (6.4–13.2) ***
Investigations	Raised inflammatory markers	0	4,583 (93.2)	20,909 (96.3)	–
		≥1	332 (6.8)	809 (3.7)	1.7 (1.3–2.1) ***
	High white cell count	0	4,664 (94.9)	21,317 (98.2)	–
		≥1	251 (5.1)	401 (1.8)	1.8 (1.3–2.5) ***
	Raised creatinine	0	4,255 (86.6)	20,050 (92.3)	–
		≥1	660 (13.4)	1,668 (7.7)	2.0 (1.6–2.4) ***
	Non-visible haematuria	0	4,603 (93.6)	21,658 (99.7)	
		≥1	312 (6.4)	60 (0.3)	28.5 (16.4–49.6) ***

*** p<0.0001(Wald test of null hypothesis that odds ratio = 1).
* p<0.05 (Wald test of null hypothesis that odds ratio = 1).

Effect modification interaction terms		**Odds ratio (95% CI)**	**p value**
Visible haematuria	× UTI		Overall: 0.0001
1	1	0.1 (0.06–0.2)	0.0001
2		0.2 (0.06–0.6)	0.005
≥3		0.2 (0.05–0.5)	0.001
1	≥2	0.04 (0.02–0.08)	0.0001
2		0.05 (0.02–0.1)	0.0001
≥3		0.2 (0.04–1.4)	0.1
Dysuria	× UTI		Overall: 0.001
≥1	1	0.4 (0.2–0.9)	0.02
	≥2	0.3 (0.1–0.6)	0.001
Non-visible haematuria	× UTI		Overall: 0.001
≥ 1	1	0.08 (0.02–0.3)	0.0001
	≥2	0.2 (0.03–0.9)	0.04

[a] *n*, Number of attendances in the analysis period.

The revised model was a significantly better fit than the model estimated using coded records alone (likelihood ratio test $\chi 2 = 1{,}546.71$, 11 d.f., $p<0.0001$). The most noteworthy result of adding text-only records to the model is that bladder cancer is strongly associated with `non-visible haematuria`, with an odds ratio of 28.5 (95% CI: 16.4–49.6) ($p<0.0001$) for one or more attendance.

The effect of adding text-only records of attendance for visible haematuria was generally unremarkable, other than to increase the odds ratio for bladder cancer in patients attending *three or more times* in the analysis period. The odds ratio for three or more attendances was estimated to be 345.7 (95% CI: 198.6–601.7) ($p<0.0001$) from coded `visible haematuria` records (see Table 10.8), compared with 644.5 (95% CI: 384.9–1,079.1) ($p<0.0001$) from the supplemented `visible haematuria` records (see Table 10.9). After supplementation with text-only records, the odds for bladder cancer significantly rose with re-attendance for visible haematuria (Wald test, $\chi 2 = 60.67$, 2 d.f., $p<0.0001$), starting at 96.6 (95% CI: 75.0–124.5) for a single visit, rising to 644.5 (95% CI: 384.9–1,079.1) for three or more.

Adding text-only records of attendance for abdominal pain had no effect on the odds ratio for bladder cancer; namely, 2.0 (95% CI: 1.7–2.5) ($p<0.001$) from `abdominal pain` records (see Table 10.8) compared with 2.0 (95% CI: 1.7–2.4) ($p<0.001$) from `abdominal pain` records (see Table 10.9).

10.2 Pancreatic cancer

10.2.1 Estimates of positive predictive value and positive likelihood ratio from coded records

The features of pancreatic cancer presentation in primary care were identified in the original study.[5] The numbers of cases and controls with coded records of attendance for each of these features are reported in Table 10.10, along with the positive likelihood ratio and positive predictive value for pancreatic cancer. The results are from my re-construction and re-analysis,[a] and largely replicated those of the original study.[5]

In terms of frequency of attendance, two features predominated: `abdominal pain` (in 1,527/3,635 = 42.0% of cases) and `jaundice` (in 1,102/3,635 = 30.3% of cases). `Jaundice` was clearly the most strongly predictive of pancreatic cancer, associated with a 12.77% (95% CI: 7.25–21.58%) chance of the disease. Indeed, patients with pancreatic cancer were 498.98 (95% CI: 268.00–929.02) times more likely to be jaundiced compared with controls. However, in all age groups combined, the positive likelihood ratio values and positive predictive values for all other features considered in isolation were unremarkable, even `abdominal pain` which was recorded quite frequently.

[a] The discrepancies were essentially trivial, and reflected very slight differences in definitions of some variables.

Table 10.10 The positive likelihood ratio and positive predictive value for pancreatic cancer in patients aged ≥40 years presenting with clinical features associated with the disease (codes))

Feature of pancreatic cancer	Number of patients attending at least once:		Positive likelihood ratio	PPV (%) (95% CI)[ab]
	Cases *n* (% of *n* = 3,635 cases)	Controls, *n* (% of 16,459 controls)	(95% CI)	
Diarrhoea	381 (10.5)	534 (3.2)	3.23 (2.85–3.67)	0.09 (0.08–0.11)
Nausea and/or vomiting	572 (15.7)	350 (2.1)	7.40 (6.51–8.41)	0.21 (0.19–0.24)
Back pain	458 (12.6)	1,028 (6.2)	2.02 (1.82–2.24)	0.06 (0.05–0.06)
Weight loss	354 (9.7)	105 (0.6)	15.27 (12.31–18.92)[c]	0.44 (0.36–0.55)
Malaise	187 (5.1)	197 (1.2)	4.30 (3.53–5.23)	0.12 (0.10–0.15)
New-onset diabetes	380 (10.5)	436 (2.7)	3.95 (3.46–4.51)	0.11 (0.10–0.13)
Jaundice	1,102 (30.3)	10 (0.06)	498.98 (268.00–929.02)[c]	12.77 (7.25–21.58)
Constipation	428 (11.8)	557 (3.4)	3.48 (3.08–3.93)	0.10 (0.09–0.11)
Abdominal pain	1,527 (42.0)	996 (6.1)	6.94 (6.46–7.45)	0.20 (0.19–0.22)

[a] PPV values are adjusted for the consulting population (see Section 6.12.3.5).
[b] PPV values estimated using Bayes' theorem (see Section 6.12.3.5), assuming a prior odds of 0.0002671 based on 2008 UK national incidence data.
[c] Positive likelihood ratio >10.

10.2.2 Effect of supplementation with text-only records on positive likelihood ratio and PPV

Detailed analysis of GPs' choice of recording style was presented in Section 9 (see Table 9.8). This section focuses on whether supplementing codes with text records alters the risk estimates for pancreatic cancer in patients with jaundice or abdominal pain. This relates to research question 5: 'Do the text data provide additional value to coded data?' (see Section 5).

The numbers of patients attending at least once for abdominal pain or jaundice determined using coded records in the original study were updated with records from the text. These revised binary abdominal pain and jaundice variables were used to re-examine the predictive power of these symptoms for pancreatic cancer (Table 10.11).

10.2.2.1 Abdominal pain

The 95% confidence intervals suggested that supplementing coded records of abdominal pain with text-only records did not alter the positive likelihood ratio (abdominal pain: 6.94, 95% CI: 6.46–7.45 vs abdominal pain: 6.59, 95% CI: 6.21–7.0) or the PPV (abdominal pain: 0.20%, 95% CI: 0.19–0.22% vs abdominal pain: 0.19%, 95% CI: 0.18–0.20%).

Table 10.11 Positive likelihood ratio and PPV for pancreatic cancer in patients of all ages presenting with abdominal pain or jaundice

Clinical feature and recording style	Number of patients attending at least once: Cases n (% of *n* = 3,635 cases)	Controls n (% of 16,459 controls)	Positive likelihood ratio (95% CI)	Positive predictive value (%) (95% CI)
Abdominal pain	1,910 (52.5)	1,312 (8.0)	6.59 (6.21–7.0)	0.19 (0.18–0.20)
Abdominal pain	1,527 (42.0)	996 (6.1)	6.94 (6.46–7.45)	0.20 (0.19–0.22)
Jaundice	1,565 (43.1)	31 (0.2)	228.59 (160.49–325.58)	6.25 (4.46–8.68)
Jaundice	1,102 (30.3)	10 (0.06)	498.98 (268.00–929.02)	12.77 (7.25–21.58)

[a] PPV values are adjusted for the consulting population (see Section 6.12.3.5). PPV values estimated using Bayes' theorem (see Section 6.12.3.5), assuming a prior odds of 0.0002671 based on 2008 UK national incidence data.

10.2.2.2 Jaundice

In contrast, supplementing the coded jaundice record with text-only *jaundice* records reduced the positive likelihood ratio from 498.98 (95% CI: 268.0–929.0) (jaundice) to 228.59 (95% CI: 160.49–325.58) (jaundice). Similarly, the PPV fell from 12.77% (95% CI: 7.25–21.58%) (jaundice) to 6.25% (95% CI: 4.46–8.68%) (jaundice).

10.2.3 Estimates of odds ratio in univariable analysis from coded records

The frequency of attendance for each feature included in the final model of the original study is reported in Table 10.12 along with its odds ratio for disease in univariable analysis. The results are from my re-construction and re-analysis, and largely match those of the original study.

10.2.3.1 Strength of association with pancreatic cancer

`Jaundice` was the feature most strongly associated with pancreatic cancer in univariable analysis – the odds of pancreatic cancer in patients who attended at least once for jaundice were 712.9 times (95% CI: 339.0–1,499.0, Wald test $p<0.0001$) higher than in non-attenders. Weight loss had the second strongest association with pancreatic cancer (odds ratio = 17.4, 95% CI: 13.8–21.9, for at least one attendance, $p<0.0001$), followed by nausea and/or vomiting (odds ratio = 9.2, 95% CI: 7.9–10.6, $p<0.0001$) and then abdominal pain (odds ratio = 7.1, 95% CI: 6.2–8.0 for a single attendance, $p<0.0001$). Diarrhoea, back pain, malaise, new-onset diabetes and constipation were all less strongly associated with pancreatic cancer, with odds ratios below 5 for at least one attendance.

Table 10.12 Frequency of attendance for features included in the final model of the original study, plus the univariable analysis

Feature of pancreatic cancer	n[a]	Number of participants		Odds ratio in univariable analysis (95% CI)	p[b]
		Cases, *n* (% of *n* = 3,635 cases)	Controls, *n* (% of *n* = 16,459 controls)		
Diarrhoea	0	3,254 (89.5)	15,925 (96.8)	–	–
	≥1	381 (10.5)	534 (3.2)	3.6 (3.1–4.1)	0.0001
Nausea and/or vomiting	0	3,063 (84.3)	16,109 (97.9)	–	–
	≥1	572 (15.7)	350 (2.1)	9.2 (7.9–10.6)	0.0001
Back pain [c]	0	3,177 (87.4)	15,431 (93.8)	–	–
	1	270 (7.4)	742 (4.5)	1.8 (1.6–2.1)	0.0001
	≥2	188 (5.2)	286 (1.7)	3.3 (2.7–4.0)	0.0001
Weight loss	0	3,281 (90.3)	16,354 (99.4)	–	–
	≥1	354 (9.7)	105 (0.6)	17.4 (13.8–21.9)	0.0001
Malaise	0	3,448 (94.9)	16,262 (98.8)	–	–
	≥1	187 (5.1)	197 (1.2)	4.8 (3.9–5.9)	0.0001
New-onset diabetes	0	3,255 (89.5)	16,023 (97.4)	–	–
	≥1	380 (10.5)	436 (2.6)	4.5 (3.9–5.3)	0.0001
Jaundice	0	2,533 (69.7)	16,449 (99.9)	–	–
	≥1	1,102 (30.3)	10 (0.06)	712.9 (339.0–1,499.0)	0.0001
Constipation	0	3,207 (88.2)	15,902 (96.6)	–	–
	≥1	428 (11.8)	557 (3.4)	4.0 (3.5–4.6)	0.0001
Abdominal pain [d]	0	2,108 (58.0)	15,463 (93.9)	–	–
	1	621 (17.1)	673 (4.1)	7.1 (6.2–8.0)	0.0001
	≥2	906 (24.9)	323 (2.0)	21.8 (18.7–25.3)	0.0001

95% CI = 95% confidence interval.
[a] n is the number of times the patient attended for the feature in the analysis period.
[b] Wald test of the null hypothesis that odds ratio is 1.

c Strong evidence that the association between `back pain` and pancreatic cancer varied with number of attendances (χ^2= 25.72, 1 d.f., p<0.0001).
d Strong evidence that the association between `abdominal pain` and pancreatic cancer varied with number of attendances (χ^2= 149.5, 1 d.f., p<0.0001).

10.2.3.2 The significance of re-attendance for a feature

There was strong evidence that re-attendance for abdominal pain was significant. The association between pancreatic cancer and `abdominal pain` varied with the number of attendances in the analysis period (χ^2= 149.5, 1 d.f., Wald test p<0.0001), with further episodes increasing the odds of disease to 21.8 (95% CI: 18.7–25.3) (p<0.0001). Similarly, the odds of pancreatic cancer increased with repeat attendance for back pain (χ^2= 25.72, 1 d.f., p<0.0001). The significance of re-attendance for diarrhoea, nausea and/or vomiting, weight loss, malaise, jaundice diabetes and constipation could not be assessed, because too few patients attended more than once to allow meaningful analysis.

10.2.4 Odds ratios in univariable analysis after supplementation with text-only records

The odds ratios for pancreatic cancer in patients attending for jaundice and abdominal pain, after supplementation with text-only records, are reported below.

10.2.4.1 Jaundice

There were high numbers of text-only *jaundice* records (Table 10.13), easily exceeding the 5% threshold for inclusion in univariable analysis. Indeed, 463 of

3,635 (12.7%) of cases attended at least once where this was never documented as a code. The text-only records indicated that patients did re-attend for jaundice; however, the numbers were too low to permit meaningful analysis in terms of association with pancreatic cancer.

Table 10.13 Frequency of presentation with jaundice in the analysis period in cases and controls, according to the recording style, and independent association with pancreatic cancer

Recording style [a]	n_a [b]	Number (%) of patients attending: Cases (*n* = 3,635)	Controls (*n* = 16,459)	Odds ratio (95% CI)	p value [c]
Jaundice	0	2,533 (6.97)	16,449 (99.9)	–	–
	≥1	1,102 (30.3)	10 (0.06)	712.9 (339.0–1,499.0)	0.0001
Jaundice	0	3,172 (87.3)	16,438 (99.9)	–	–
	≥1	463 (12.7)	21 (0.1)	130.3 (79.2–214.6)	0.0001
Jaundice	0	2,070 (56.7)	16,428 (99.8)	–	–
	≥1	1,565 (43.1)	31 (0.2)	640.4 (353.9–1,158.9)	0.0001

[a] Strong evidence that the association between pancreatic cancer and jaundice varied with the recording style ($\chi 2 = 7.63$, 1 d.f., $p<0.01$).

[b] Number of attendances in the analysis period (the maximum number of attendances examined was that level containing no fewer than 5% of cases).

[c] Wald test of the null hypothesis that the odds ratio is 1.

There was strong evidence that the association between jaundice and pancreatic cancer varied with the recording style (Wald test, $\chi 2 = 7.63$, 1 d.f., $p<0.01$). The odds ratio for pancreatic cancer from coded (jaundice) records was 1,132.4 (95% CI: 506.1–2,533.8) (Wald test, $p<0.0001$) for at least one

attendance, whereas it was 314.4 (95% CI: 156.5–631.4) (p<0.0001) according to the text-only (*jaundice*) records.

The odds ratio for the combined variable `jaundice` was 640.4 (95% CI: 353.9–1,158.9) for at least one attendance. The numbers re-attending with jaundice were too low to permit meaningful analysis of its significance in terms of altered association with pancreatic cancer.

10.2.4.2 Abdominal pain

There was a significant amount of text-only *abdominal pain* recording: 383 of 3,635 cases (10.5%) attended at least once, meeting the 5% threshold for inclusion in univariable analysis (see Table 10.14). There was strong evidence that the association between pancreatic cancer and abdominal pain varied with the recording style ($\chi 2$= 151.44, 2 d.f., p<0.0001), although the size of the difference was of questionable clinical significance. The odds ratio for pancreatic cancer from text-only *abdominal pain* records was 6.5 (95% CI: 5.5–7.6) (Wald test, p<0.0001) for at least one attendance, whereas it was 7.1 (95% CI: 6.2–8.0) (p<0.0001) for a single attendance according to the coded `abdominal pain` records.

The odds ratio for the combined variable `abdominal pain` was 7.3 (95% CI: 6.5–8.2, p<0.0001) for a single attendance. The effect of re-attendance for abdominal pain in the analysis period was highly significant ($\chi 2$= 229.31, 1 d.f., p<0.0001), with the odds ratio rising to 26.3 (95% CI: 22.8–30.3) (p<0.0001) for two or more attendances.

Table 10.14 Frequency of attendance for abdominal pain in the analysis period in cases and controls grouped by recording style, and independent association with pancreatic cancer

Recording style[a]	n_a[b]	Number (%) of patients attending for abdominal pain		Odds ratio (95% CI)	p value[c]
		Cases (*n* = 3,635)	Controls (*n* = 16,459)		
Abdominal pain	0	2,108 (58.0)	15,463 (94.0)	–	–
	1	621 (17.1)	673 (4.1)	7.1 (6.2–8.0)	0.0001
	≥2	906 (24.9)	323 (2.0)	21.8 (18.7–25.3)	0.0001
Abdominal pain	0	3,252 (89.5)	16,143 (98.1)	–	–
	≥1	383 (10.5)	316 (1.9)	6.5 (5.5–7.6)	0.0001
Abdominal pain	0	1,725 (47.5)	15,147 (92.0)	–	–
	1	734 (20.2)	891 (5.4)	7.3 (6.5–8.2)	0.0001
	≥2	1,176 (32.4)	421 (2.6)	26.3 (22.8–30.3)	0.0001

[a] Strong evidence that the association between abdominal pain and pancreatic cancer varied with recording style ($\chi 2$= 151.44, 2 d.f., p<0.0001).
[b] Number of attendances in the analysis period.
[c] Wald test that the odds ratio is 1.

10.2.5 Multivariable analysis

10.2.5.1 The basic model estimated using coded records

The multivariable conditional logistic regression analysis of the presenting features of pancreatic cancer, as identified in the original study,[5] was run using solely code-based variables to replicate the basic model (Table 10.15). Jaundice was the feature most strongly associated with pancreatic cancer. The odds of pancreatic cancer were 1,306.1 (95% CI: 599.9–2,843.7) (Wald

test, p<0.0001) times higher in patients who attended at least once with this feature compared with non-attenders. The significance of re-attendance for jaundice could not be assessed, owing to small numbers of controls.

`Abdominal pain` was also associated with pancreatic cancer, albeit less strongly than `jaundice`. The odds of disease in patients attending once with abdominal pain were 7.9 (95% CI: 6.6–9.5) (Wald test, p<0.0001). There was strong evidence that the association between `abdominal pain` and pancreatic cancer varied with repeat attendance ($\chi 2$= 78.18, 1 d.f., p<0.0001). The odds of disease increased from 7.9 (95% CI: 6.6–9.5) for a single visit to 23.4 (95% CI: 18.9–28.8) for two or more.

Also strongly associated with pancreatic cancer in multivariable analysis were `weight loss` (odds ratio 19.3, 95% CI: 13.7–27.1, p<0.0001) and `nausea and/or vomiting` (6.2, 95% CI: 4.9–7.8, p<0.0001). The remaining features were less strongly associated with pancreatic cancer (odds ratio lower than 5).

Two or more attendances for back pain were more significant than a single episode ($\chi 2$= 8.42 1 d.f., p<0.01).

Table 10.15 Frequency of attendance for symptoms of pancreatic cancer in the year before analysis and odds ratio in multivariable conditional logistic regression analysis; estimates were obtained from code-based variables

Feature of pancreatic cancer	n[a]	Number of:		Odds ratio (95% CI)
		Cases (*n* = 3,635)	Controls (*n* = 16,459)	
Diarrhoea	0	3,254 (89.5)	15,925 (96.8)	–
	≥1	381 (10.5)	534 (3.2)	2.4 (1.9–3.1)***
Nausea and/or vomiting	0	3,063 (84.3)	16,109 (97.9)	–
	≥1	572 (15.7)	350 (2.1)	6.2 (4.9–7.8)***
Back pain	0	3,177 (87.4)	15,431 (93.8)	–
	1	270 (7.4)	742 (4.5)	1.3 (1.1–1.7)*
	≥2	188 (5.2)	286 (1.7)	2.4 (1.7–3.3)***
Weight loss	0	3,281 (90.3)	16,354 (99.4)	–
	≥1	354 (9.7)	105 (0.6)	19.3 (13.7–27.1)***
Malaise	0	3,448 (94.9)	16,262 (98.8)	–
	≥1	187 (5.1)	197 (1.2)	2.9 (2.1–4.1)***
New-onset diabetes	0	3,255 (89.5)	16,023 (97.4)	–
	≥1	380 (10.5)	436 (2.7)	4.2 (3.4–5.4)***
Jaundice	0	2,533 (69.7)	16,449 (99.9)	–
	≥1	1,102 (30.3)	10 (0.06)	1,306.1 (599.9–2,843.7)***
Constipation	0	3,207 (88.2)	15,902 (96.6)	–
	≥1	428 (11.8)	557 (3.4)	2.1 (1.7–2.7)***

Feature of pancreatic cancer	n[a]	Number of: Cases (*n* = 3,635)	Controls (*n* = 16,459)	Odds ratio (95% CI)
Abdominal pain	0	2,108 (58.0)	15,463 (94.0)	–
	1	621 (17.1)	673 (4.1)	7.9 (6.6–9.5)***
	≥2	906 (24.9)	323 (2.0)	23.4 (18.9–28.8)***

[a]Number of attendances in the analysis period.

**p<0.01, ***p<0.0001, Wald test of the null hypothesis that the odds ratio is 1.

10.2.5.2 Goodness of fit with and without text-based variables

The effect of adding the new, text-based variables *abdominal pain* and *jaundice* was highly significant (nested models, likelihood ratio $\chi 2$ = 1,246.10, 2 d.f., p<0.0001). Therefore, text-only records were included for both features in the revised final model (see Table 10.16).

10.2.5.3 Effect modification

There was significant and antagonistic effect modification between abdominal pain and jaundice (Wald test, $\chi 2$ = 22.1, 2 d.f., p<0.0001). In addition, there was also significant and antagonistic effect modification between constipation and abdominal pain (Wald test, $\chi 2$ = 21.51, 2 d.f., p<0.0001). The individual interaction terms for effect modification are reported in Table 10.16.

10.2.6 Revised final model

The revised final model, including the text-only records and effect modification, is reported in Table 10.16. The revised model was a significantly better fit than the model estimated using coded records alone (likelihood ratio text $\chi 2$ = 1,322.95, 4 d.f., p<0.0001).

The addition of text-only records for jaundice and abdominal pain did not appreciably alter the odds ratios for these features. Jaundice remained the most strongly associated with pancreatic cancer, whether the odds ratio was estimated using coded `jaundice` records (main effects odds ratio: 1,306.1, 95% CI: 599.9–2,843.7, Wald test, p<0.0001) (see Table 10.15) or the supplemented `jaundice` records (1,969.7, 95% CI: 918.2–4,225.3, p<0.0001) records (see Table 10.16). While the point estimate from the `jaundice` records appears larger than that from the `jaundice` records, the wide confidence intervals for both estimates suggests that there is no difference between them.

Similarly, supplementing with text records of abdominal pain attendance did not appreciably alter the association between this symptom and pancreatic cancer. The odds ratio was estimated to be 7.9 (95% CI: 6.6–9.5, p<0.0001) for a single attendance from coded `abdominal pain` records (see Table 10.15) compared with 7.2 (95% CI: 5.9–8.8, p<0.0001) from `abdominal pain` records (see Table 10.16).

Re-attendance for abdominal pain was significant (χ^2= 127.54, 1 d.f., p<0.0001) with the odds ratio rising to 35.6 (95% CI: 27.9–45.3) for two or more attendances (see Table 10.16 on the next page).

In the final model, the significance of a single attendance for back pain was lost, the symptom only becoming significant upon the second attendance, with an odds ratio of 2.6 (95% CI: 1.8–3.7) (p<0.0001).

The odds ratios for the rest of the features of pancreatic cancer were only adjusted slightly by the addition of text-only records for jaundice and abdominal pain.

Table 10.16 Conditional logistic regression analysis of the final model of pancreatic cancer presentation

Feature of pancreatic cancer	n[a]	Number of: Cases (*n* = 3,635)	Controls (*n* = 16,459)	Odds ratio (95% CI)
Main effects				
Diarrhoea	0	3,254 (89.5)	15,925 (96.8)	–
	≥1	381 (10.5)	534 (3.2)	1.5 (1.1–2.0)*
Nausea and/or vomiting	0	3,063 (84.3)	16,109 (97.9)	–
	≥1	572 (15.7)	350 (2.1)	5.3 (4.1–7.0)***
Back pain	0	3,177 (87.4)	15,431 (93.8)	–
	1	270 (7.4)	742 (4.5)	1.3 (1.0–1.7)†
	≥2	188 (5.2)	286 (1.7)	2.6 (1.8–3.7)***
Weight loss	0	3,281 (90.3)	16,354 (99.4)	–
	≥1	354 (9.7)	105 (0.6)	27.5 (18.2–41.5)***
Malaise	0	3,448 (94.9)	16,262 (98.8)	–
	≥1	187 (5.1)	197 (1.2)	2.3 (1.6–3.5)***
New-onset diabetes	0	3,255 (89.5)	16,023 (97.4)	–
	≥1	380 (10.5)	436 (2.7)	4.6 (3.5–6.0)***
Jaundice	0	2,070 (57.0)	16,428 (99.8)	–
	≥1	1,565 (43.0)	31 (0.2)	1,969.7 (918.2–4,225.3)***
Constipation	0	3,207 (88.2)	15,902 (96.6)	–

	≥1	428 (11.8)	557 (3.4)	3.0 (2.2–4.2)***
Abdominal pain	0	1,725 (47.5)	15,147 (92.0)	–
	1	734 (20.2)	891 (5.4)	7.2 (5.9–8.8)***
	≥2	1,176 (32.4)	421 (2.6)	35.6 (27.9–45.3)***
Effect modification interaction terms			**Odds ratio (95% CI)**	**p value**
Abdominal pain	×	Jaundice		0.001
1		1	0.1 (0.05–0.5)	
≥2		1	0.1 (0.05–0.2)	
Abdominal pain	×	Constipation		0.0001
1		1	0.4 (0.2–0.8)	
≥2		1	0.3 (0.2–0.5)	

[a] Number of attendances in the analysis period.

†p = 0.06, *p<0.01, ***p<0.0001 Wald test that odds ratio is 1.

11 Results: Comparison of diagnostic intervals estimated from coded and from text-only records

The diagnostic interval is estimated as the number of days between first attendance for a symptom and the date of cancer diagnosis. This analysis of diagnostic intervals estimated from coded and text-only records of attendance for symptoms relates to research question 5: Do the text data provide additional value to coded data? The methods are described in Section 6.12.2.2.

11.1.1 Summary statistics for diagnostic interval data

The change in diagnostic interval following the addition of text-only records of attendance for haematuria, abdominal pain and jaundice is reported in Table 11.1. As explained in Section 6.12.2.2, there are no suitable methods for testing whether there is a statistical difference between the diagnostic interval as estimated from the earliest coded record (DI_{coded}) and that estimated from the earliest ever record (whether recorded as a code or in the text, $DI_{text/coded}$). This is because the data are neither fully matched nor completely unmatched. The wide interquartile range suggests that there is no difference in the point estimates of diagnostic interval regardless of the recording method used.

Table 11.1 Diagnostic interval data (median, 25% to 75% interquartile range, IQR) estimated before and after addition of text-only records of attendance for haematuria, abdominal pain and jaundice

Cancer site	Symptom	n_{coded}	DI_{coded}, days (25% to 75% IQR)	$n_{text/coded}$	$DI_{text/coded}$, days (25% to 75% IQR)
Bladder	Visible haematuria	2,595	−65 (−122 to −35)	3,147	−69 (−130 to −36)
	Non−visible haematuria	127	−77 (−131 to −41)	312	−65 (−114 to −33)
	Abdominal pain	358	−134 (−249 to −54)	547	−123 (−244 to −47)
Pancreas	Jaundice	1,110	−27 (−50 to −13)	1,565	−25 (−49 to −12)
	Abdominal pain	1,527	−85 (−175 to −42)	1,910	−78 (−172 to −37)

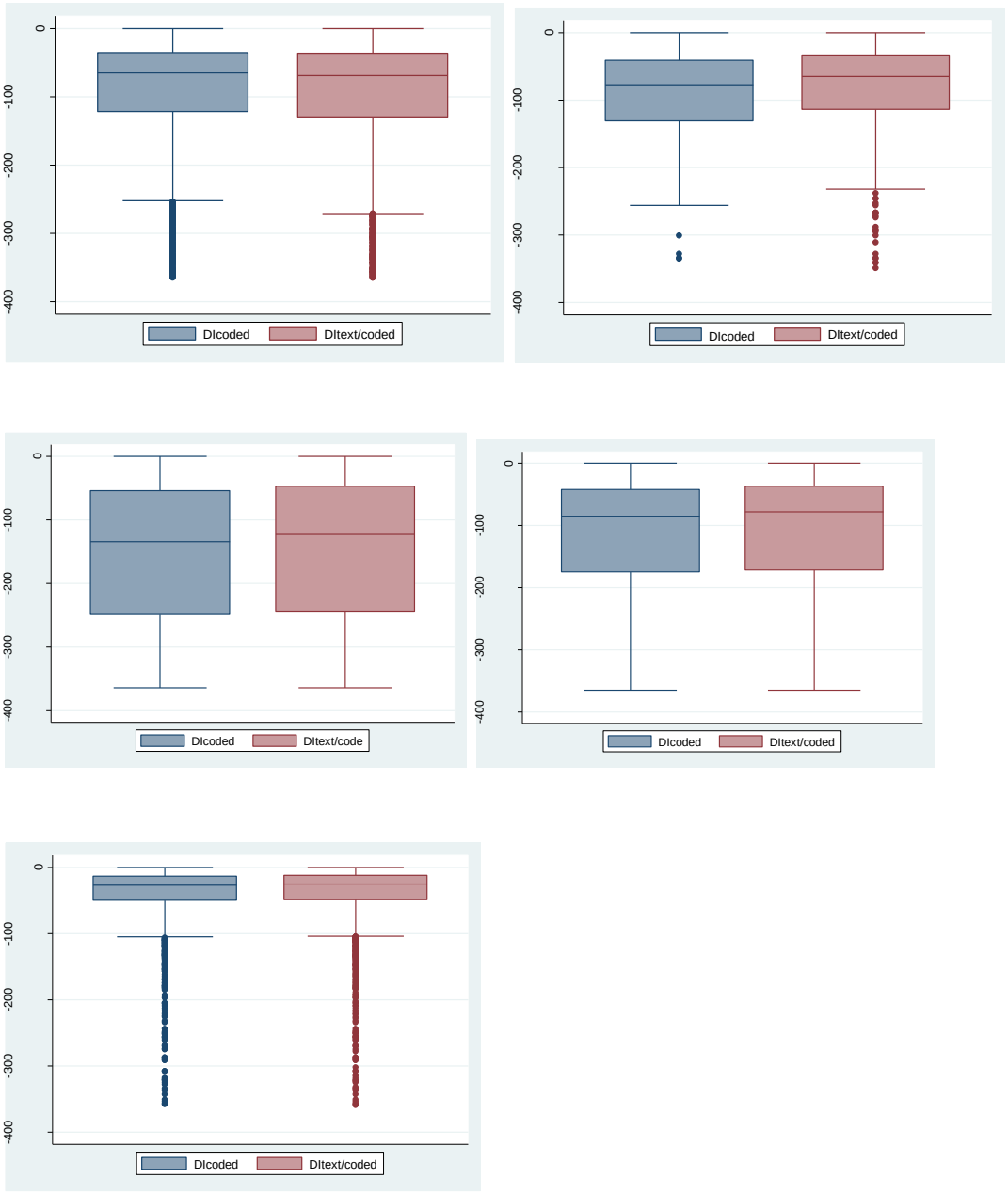
Note: n_{coded}: number of patients identified from coded records as having attended at least once for the symptom in the analysis period.

$n_{text/coded}$: number of patients identified from coded or text records as having attended at least once for the symptom in the analysis period.

11.1.1.1 Testing for normality

None of the diagnostic interval data reported in Table 11.1 followed a normal distribution (D'Agostino K-squared test of normality, $p<0.001$ for each symptom before or after addition of text-only records). The box-and-whisker plots are shown in Figure 11.1. Therefore, non-parametric methods were used in the matched and unmatched analysis reported below.

Figure 11.1 Box and whisker plot of diagnostic interval for visible (top left), non-visible (top right) haematuria and abdominal pain (middle left) in bladder cancer and for abdominal pain (middle right) and jaundice (bottom) in pancreatic cancer. Estimates were made using the first coded record (DIcoded, blue) and the first ever record (DItext/code, pink)



11.1.1.2 Matched analysis

The results of matched analysis are reported in Table 11.2. Patients were included in the analysis if they had both coded and text records of attendance for a symptom. For each patient the diagnostic interval was estimated from the earliest coded record (DI_{coded}) and from the earliest record, which may have been coded or in the text ($DI_{text/coded}$).

There was strong evidence that the diagnostic interval for each symptom was lengthened by the addition of text-only records of attendance.[a] This was most marked for abdominal pain in the bladder cancer dataset, where the coded records estimated diagnostic interval as −130.5 (25% to 75% interquartile range: −249 to −68) days compared with −147.5 (25% to 75% interquartile range: −260 to −81) as estimated after addition of text-only records. The change in diagnostic interval was smallest for jaundice in pancreatic cancer, where the difference in median estimates of diagnostic interval was only 3 days.

[a] It would be impossible for addition of text-only records to shorten diagnostic intervals.

Table 11.2 Matched analysis of diagnostic interval data – patients who had both a coded and a text record of attendance

Site	Symptom	$n_{coded+text}$	DI_{coded}, days (25% to 75% IQR)	$DI_{text/coded}$ (25% to 75% IQR)	p[a]
Bladder	Visible haematuria	1,438	−67.5 (−119 to −35)	−77 (−137 to −41)	<0.0001
	Non-visible haematuria	30	−76.5 (−128 to −41)	−82.5 (−135 to −55)	<0.01
	Abdominal pain	138	−130.5 (−249 to −68)	−147.5 (−260 to −81)	<0.0001
Pancreas	Jaundice	695	−28 (−53 to −14)	−31 (−59 to −16)	<0.0001
	Abdominal pain	739	−95 (−196 to −48)	−105 (−222 to −54)	<0.0001

Notes:

[a] Sign rank test of the null hypothesis $DI_{coded} - DI_{text/coded} = 0$ in matched pairs.

$n_{coded+text}$: number of bladder or pancreatic cancer cases with a coded and a text record of the symptom.

DI_{coded}: number of days from the earliest coded record of the symptom and diagnosis with bladder or pancreatic cancer.

$DI_{text/coded}$: number of days from the earliest text or coded record of the symptom and diagnosis with bladder or pancreatic cancer.

11.1.1.3 Unmatched analysis

The results of unmatched analysis are reported in Table 11.3. Apart from visible haematuria in the bladder cancer dataset, the results suggest that diagnostic intervals from coded (DI_{coded}) and from text (DI_{text}) records are *not* drawn from the same population (Wilcoxon rank-sum test, $p<0.01$ to $p<0.0001$).

The unmatched data analysis suggests that the diagnostic interval estimated from the first coded record is greater than that estimated from the first text-only record. This contrasts with the results in matched analysis, where addition of text-only records lengthened the diagnostic interval.

The difference in median values of diagnostic interval was greatest for abdominal pain in bladder cancer; namely, −136.5 (25% to 75% interquartile range: −249 to −48) days (coded records) and −90 (−210 to −34) days (text-only records). The difference in median values was least for jaundice in pancreatic cancer, i.e. −23 (25% to 75% interquartile range: −47 to −11) days (coded records) and −19 (−36 to −7) days (text-only records).

Table 11.3 Unmatched analysis of diagnostic interval data

Site	Symptom	n_{coded}	n_{text}	DI_{coded}, days (25% to 75% IQR)	DI_{text} (25% to 75% IQR)	p[a]
Bladder	Visible haematuria	1,157	552	−63 (−125 to −35)	−60 (−111.5 to −29)	0.07
	Non-visible haematuria	97	185	−77 (−132 to −42)	−56 (−100 to −30)	0.015
	Abdominal pain	220	189	−136.5 (−249 to −48)	−90 (−210 to −34)	0.018
Pancreas	Jaundice	415	455	−23 (−47 to −11)	−19 (−36 to −7)	0.0035
	Abdominal pain	788	383	−76 (−163 to −36)	−46 (−106 to −19)	<0.0001

Notes:

[a] Wilcoxon rank-sum (Mann-Whitney) test of the null hypothesis that DI_{coded} and DI_{text} are drawn from the same population.

n_{coded}: number of bladder or pancreatic cancer cases with solely a coded record of the symptom.

n_{text}: number of bladder or pancreatic cancer cases with solely a text record of the symptom.

12 Results: Modelling the outcome ‘text-only recording’ of visible haematuria in the bladder cancer dataset

This *post-hoc* analysis was conducted to explore the hypothesis that GPs tend to use codes to record clinically significant events and text for anything perceived to be less worrisome. This also relates to research question 4: Does the recording style vary with the clinical context of presentation of a symptom? The methods used are described in Section 6.12.5.

12.1 Relationship between recording style and cause of visible haematuria

For those patients who attended at least once for visible haematuria in the analysis period, the recording style used was tabulated against the possible cause; namely, bladder cancer and urinary tract infection (Table 12.1, with graphical presentation in Figure 12.1).[a] [b] There were insufficient numbers of patients with calculi to allow meaningful analysis (*n* = 12).

[a] Note that text-only recording status was reserved for just those patients whose visible haematuria was only ever documented in the text.

[b] As explained in Methods, patient case–control status was used as a proxy for malignant versus benign cause.

Table 12.1 Recording style of visible haematuria tabulated against possible cause and patient gender

Explanatory variable		No. (%) of patients attending at least once estimated using: [a]		
		Visible haematuria	*Visible haematuria*	**Visible haematuria**
Benign cause vs bladder cancer [b]	Control	196 (58.3)	140 (41.7)	336 (100)
	Case	2,591 (82.3)	556 (17.7)	3,147 (100)
	Total	*2,787 (80.0)*	*696 (20.0)*	*3,483 (100)*
Urinary tract infection [c, d]	Attended ≥1	450 (72.0)	175 (28.0)	625 (100)
	No attendance	2,337 (81.8)	521 (18.2)	2,858 (100)
	Total	*2,787 (80.0)*	*696 (20.0)*	*3,483 (100)*
Gender [e]	Male	2,132 (80.9)	503 (19.1)	2,635 (100)
	Female	655 (77.2)	193 (22.8)	848 (100)
	Total	*2,787 (80.0)*	*696 (20.0)*	*3,483 (100)*

[a] Visible haematuria – GP used a Read code for some or all attendances; *Visible haematuria* – GPs used text-only to record all attendances; Visible haematuria – total record.

[b] Strong evidence against the null hypothesis that there is no relationship between recording style and benign vs malignant cause of haematuria (Pearson $\chi 2$, 1 d.f., = 109.4, $\chi 2$ test: $p<0.0001$).

[c] Strong evidence against the null hypothesis that there is no relationship between recording style and attendance for urinary tract infection (Pearson $\chi 2$, 1 d.f., = 30.6, $\chi 2$ test: $p<0.0001$).

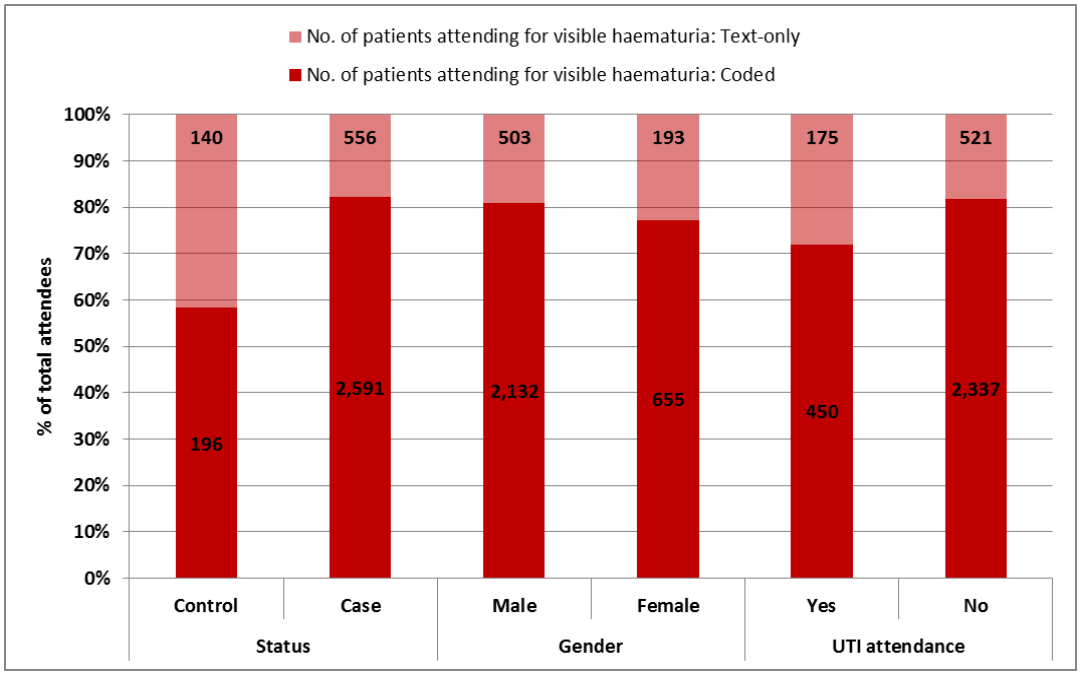
[d] Occurrences of urinary tract infection and visible haematuria were not matched on date.

[e] Moderate evidence against the null hypothesis that there is no relationship between recording style and gender (Pearson $\chi 2$, 1 d.f., = 5.4, $\chi 2$ test: $p<0.05$).

As reported above (see Section 9.2.2.1), recording style varied with patient case/control status (χ2 test: p<0.0001). The preference for text-only recording was greater in controls (140/336 = 41.7%) – whose haematuria is assumed to have a benign cause – than in patients diagnosed with bladder cancer within the year (i.e. the cases, 556/3,147 = 17.7%) (Table 12.1).

Recording style also varied according to whether patients had attended for a urinary tract infection (χ2 test: p<0.0001): text-only recording was more highly favoured in patients with a history of urinary tract infection (175/625 = 28.0%) than in those who had no record of attendance for this (521/2,858 = 18.2%) (Table 12.1).

Figure 12.1 Recording style of visible haematuria plotted as a function of possible cause and patient gender



12.2 Relationship between recording style and gender

There was moderate evidence ($\chi 2$ test: $p<0.05$) of a small variation of recording style with gender (cases and controls combined), with the GP preference for text-only recording being slightly more marked in women (193/848 = 22.8%) than in men (503/2,635 = 19.1%) (Table 12.1 and Figure 12.1). The results of earlier analysis (see Figure 9.6) show that this is particularly marked for female controls, where the records of attendance for visible haematuria were lost in the hidden text for 38 of the 57 (66.7%) female controls.

12.3 Univariable analysis

In univariable analyses, benign vs malignant cause (i.e. control vs case status) ($p<0.001$), attendance for a urinary tract infection ($p<0.001$) and gender ($p<0.05$) were all independently associated with the outcome 'text-only recording' (mixed-effects logistic regression) (Table 12.2).

Table 12.2 Univariable analysis for the outcome 'text-only recording of visible haematuria' (mixed-effects logistic regression, controlling for random effects of clustering within GP practice)

Explanatory variable	Odds ratio (univariable analysis)	95% CI	p value [d]
Benign vs malignant cause (control vs case)[a]	3.6	2.8–4.7	0.0001
Urinary tract infection (≥1 attendance vs no attendance) [b]	1.9	1.5–2.4	0.0001
Gender (females vs males) [c]	1.3	1.0–1.5	0.02

[a] Random effect of clustering within GP practice was significant (χ^2= 29.9, p<0.0001).

[b] Random effect of clustering within GP practice was significant (χ^2= 32.8, p<0.0001).

[c] Random effect of clustering within GP practice was significant (χ^2= 27.9, p<0.0001).

[d] z-test that the odds ratio is not 1.

12.4 Modification of the effect of benign vs malignant causes by gender

Initial analysis involved looking for evidence of an association between benign versus malignant causes and text-only recording within gender. For men, the preference for text-only recording of haematuria was greater in the controls (presumed to have a benign cause) (102/279 = 36.6%) than in the cases (presumed to be due to bladder cancer) (401/2,356 = 17.0%) (χ^2 test, p<0.0001). The preference for text-only recording in controls was more marked in the women (38/57 = 66.7% in controls vs 155/791 = 19.6% in cases) (χ^2 test, p<0.0001) (see Table 12.3 and Figure 12.2).

These results justified testing, in the final model, whether the association between the suspected cause of haematuria (benign versus malignant) and its recording solely in the text was different in male compared with female patients.

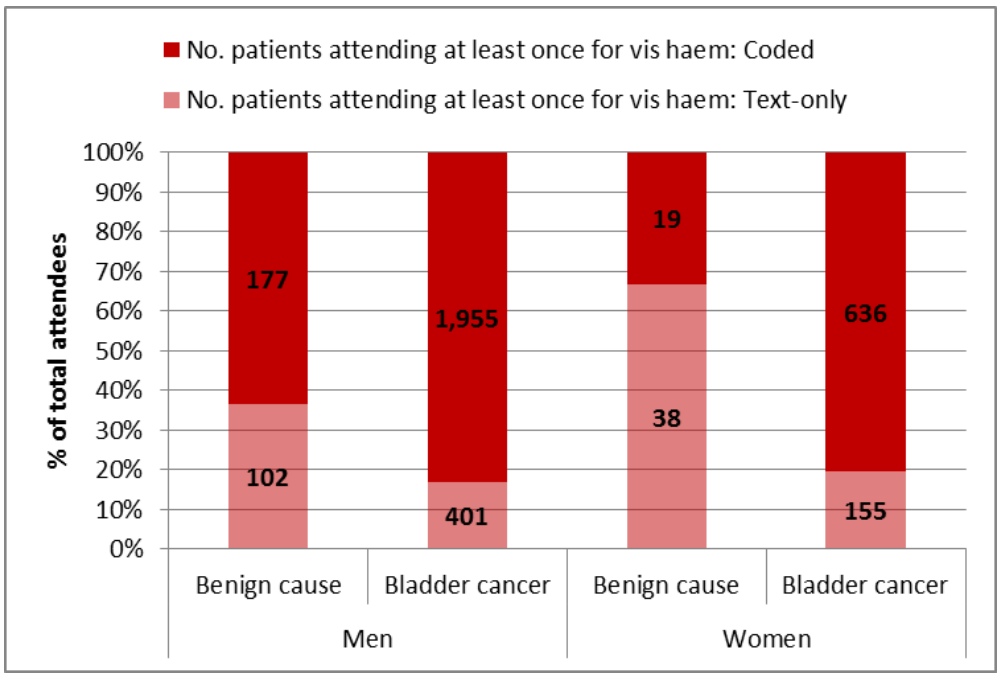
Table 12.3 Overall numbers of patients in the bladder cancer dataset presenting at least once with visible haematuria recorded as a Read code or solely in the text, grouped by cause of visible haematuria within gender

Gender	Benign cause vs bladder cancer	Number of patients (%) attending at least once estimated using:		
		Visible haematuria	*Visible haematuria*	**Visible haematuria**
Male [a]	**Benign (controls)**	177 (63.4)	102 (36.6)	279 (100)
	Bladder cancer (cases)	1,955 (83.0)	401 (17.0)	2,356 (100)
	Subtotal	*2,132 (80.9)*	*503 (19.1)*	*2,635 (100)*
Female [b]	**Benign (controls)**	19 (33.3)	38 (66.7)	57 (100)
	Bladder cancer (cases)	636 (80.4)	155 (19.6)	791 (100)
	Subtotal	*655 (77.2)*	*193 (22.8)*	*848 (100)*
Total		2,787 (80.0)	696 (20.0)	3,483 (100)

[a] $\chi 2$ test: $p<0.0001$ – strong evidence to reject the null hypothesis that, within men, there is no association between benign vs malignant cause of visible haematuria and recording style.

[b] $\chi 2$ test: $p<0.0001$ – strong evidence to reject the null hypothesis that, within women, there is no association between benign vs malignant cause of visible haematuria and recording style.

Figure 12.2 Overall numbers of patients in the bladder cancer dataset presenting at least once with visible haematuria recorded as a Read code or solely in the text, grouped by gender within cause of visible haematuria



12.5 Modification of the effect of urinary tract infection by gender

For male patients there was a greater tendency for their attendance for visible haematuria to be recorded solely in the text if they had a history of a urinary tract infection (107/384 = 27.9%) than if they did not (396/2,251 = 17.6%). A similar pattern was observed in the women (68/241 = 28.2% in women with a history of a urinary tract infection, compared with 125/607 = 20.6% in women with no such history) (Table 12.4 and Figure 12.3). Indeed, the results

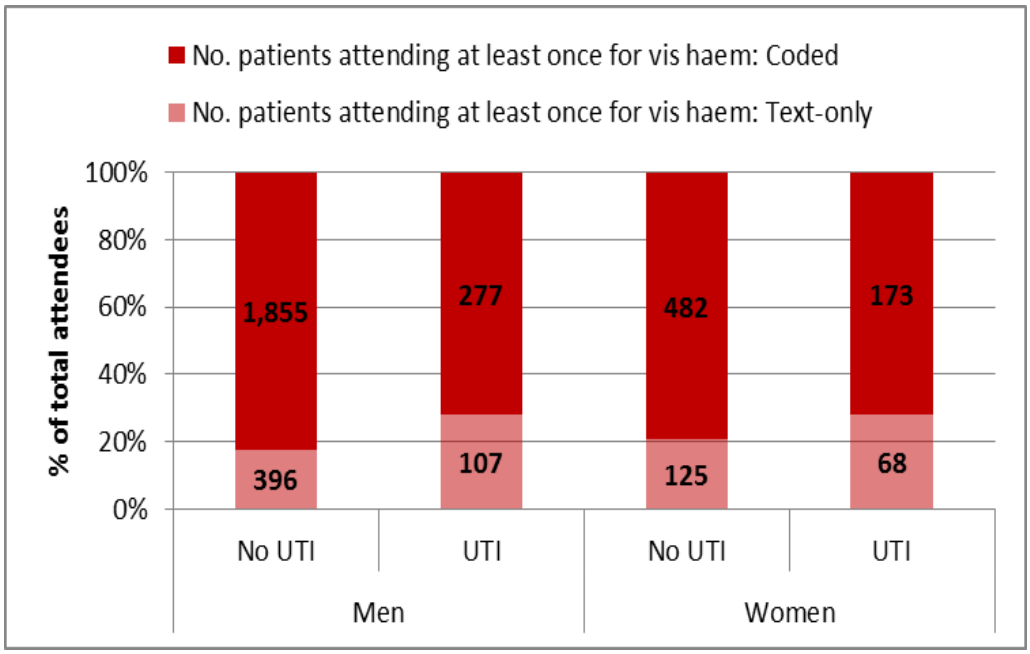
suggested an association between recording style and a history of urinary tract infection in men ($\chi 2$ test, $p<0.0001$) and women ($\chi 2$ test, $p<0.05$).

These results justified testing, in the final model, for modification of the effect of urinary tract infection by gender.

Table 12.4 Overall numbers of patients in the bladder cancer dataset presenting at least once with visible haematuria recorded as a Read code or as text only, grouped by a history of urinary tract infection (UTI) within gender

Gender	UTI	Number of patients (%) attending at least once estimated using:		
		Visible haematuria	*Visible haematuria*	Visible haematuria
Male	No UTI	1,855 (82.4)	396 (17.6)	2,251 (100)
	UTI	277 (72.1)	107 (27.9)	384 (100)
	Subtotal	*2,132 (80.9)*	*503 (19.1)*	*2,635 (100)*
Female	No UTI	482 (79.4)	125 (20.6)	607 (100)
	UTI	173 (71.8)	68 (28.2)	241 (100)
	Subtotal	*655 (77.4)*	*193 (22.6)*	*848 (100)*
Total		**2,787 (80.0)**	**696 (20.0)**	**3,483 (100)**

Figure 12.3 Overall numbers of patients in the bladder cancer dataset presenting at least once with visible haematuria recorded as a Read code or as text only, grouped by history of urinary tract infection (UTI) within gender



12.6 Multivariable analysis – main effects

The main effects of patient case/control status,[a] attendance for urinary tract infection and gender were significant in multivariable analysis (see Table 12.5). Patient case/control status was most strongly associated with the outcome 'text-only recording of visible haematuria' (odds ratio 3.7, 95% CI: 2.8–4.7, p<0.0001), followed by urinary tract infection (odds ratio 1.9, 95% CI: 1.8–2.2, p<0.0001). Women were more likely than men to have their haematuria recorded in the text (odds ratio: 1.2, 95% CI: 1.0–1.5, p<0.05). Clustering within

[a] Note this is a proxy for cause of haematuria, i.e. assumed to be benign in the controls and bladder cancer in the cases.

GP practice was significant and adjustment for this was retained (likelihood ratio test, p<0.001).

Table 12.5 Multivariable analysis for the outcome 'text-only recording of visible haematuria' (mixed-effects logistic regression, controlling for random effects of clustering within GP practice)

Explanatory variable	Odds ratio (multivariable analysis) [a]	95% CI	p value [b]
Control–case status (benign cause vs bladder cancer)	3.7	2.8–4.7	0.0001
Gender (female vs male)	1.2	1.0–1.5	0.04
Urinary tract infection (≥1 attendance vs no attendance)	1.8	1.4–2.2	0.0001
Constant	0.2	0.1–0.2	0.0001

[a] Random effect of clustering within GP practice was significant ($\chi 2$= 34.7, p<0.0001).

[b] z-test of the null hypothesis that the odds ratio is 1.

12.6.1 Effect modification

There was no evidence that the effect of a history of urinary tract infection on text-only recording of visible haematuria was different in men and women (interaction term not significant, p = 0.4).

In contrast, the effect of benign versus malignant cause of visible haematuria on the style used by GPs to record the symptom was markedly different in men and women (interaction term 3.1, 95% CI: 1.6–6.2, p<0.001); therefore, this was retained in the final model (Table 12.6).

12.6.2 The final model

The final model is reported in Table 12.6. Suspicion of an underlying benign pathology was positively associated with text-only recording of visible haematuria. This was most strongly observed in female controls attending for visible haematuria, for whom the odds ratio of text-only recording was 9.3 (95% CI: 5.0–17.4, p<0.0001). In male controls, the odds ratio was less, at 3.0 (95% CI: 2.2–4.0, p<0.0001). For this reason, the main effect of gender was retained in the final model, despite the fact that it was not significant (p=0.4). The odds that GPs would record attendance for visible haematuria solely in the text were 1.8 times (95% CI: 1.4–2.3, p<0.0001) higher if patients had a history of urinary tract infection. The random effect of clustering within GP practice (the estimated standard deviation in the intercept on the logit scale is 0.4, 95% CI: 0.3–0.7) was significant ($\chi 2$ = 34.8, p<0.0001).

Table 12.6 Multivariable analysis for the outcome 'text-only recording of visible haematuria' (mixed-effects logistic regression, controlling for random effects of clustering within GP practice)

Explanatory variable	Odds ratio (multivariable analysis)	95% Confidence interval	P value[a]
Benign cause vs bladder cancer (in men)	3.0	2.2–4.0	0.0001
Benign cause vs bladder cancer (in women)	9.3	5.0–17.4	0.0001
UTI vs no UTI	1.8	1.4–2.3	0.0001
Female vs male	1.1	0.9–1.4	0.4
Constant	0.2	0.1–0.2	0.0001

Note to Table 12.6:

[a] z-test that the odds ratio is not 1.

13 Discussion

This was primarily a methodological study, albeit with some clinical implications, using the Clinical Practice Research Datalink (CPRD) primary care database. The methodological objectives were devised to investigate recently developed research methods for cancer diagnosis that use electronic general practice databases; for example, methods underpinning the derivation of QCancer[7] and risk assessment tools.[4,5] This is important, because many studies using these techniques informed recently updated guidance from the National Institute for Health and Care Excellence (NICE) on cancer referral from primary care. Indeed, the recommendations in these guidelines influence NHS spending approaching £1bn annually.[9]

To carry out the study, two datasets used to characterise the presentation of bladder[4] and pancreatic[5] cancer in primary care were re-created and then augmented with previously unavailable, 'hidden', text records of attendance for haematuria, jaundice and abdominal pain. This newly acquired information allowed me to identify – and, to an extent, quantify – methodological problems inherent in symptom-based studies that use such databases but omit text records from analysis.

This chapter starts with a discussion of the validity of the methods used in this thesis (Section 13.1), as well as their strengths and limitations (Section 13.2). Discussion of my study's methodological findings follows next (see Section 13.3), including an outline of their implications for future research using data

from the CPRD. The impact of the methodological findings on clinical outcome measures is discussed in Section 13.4. Finally, the new clinical findings of this study are discussed in Section 13.5.

13.1 Choice of study design

To a large extent, the study design was governed by decisions made by researchers who planned the original studies that were extended by my PhD.[4,5] The case–control designs, set in the CPRD, were chosen because they permitted the study of prospective medical records of a sufficiently large number of patients diagnosed with the cancers of interest. The study design also enabled multiple symptoms and signs reported by the patients to be investigated.

My study design has a number of strengths and limitations, which are discussed below.

13.2 Study strengths and limitations

13.2.1 Setting

A shared aim of the original studies was to improve the selection of primary care patients referred to secondary care for investigation of possible cancer. Therefore, the CPRD – generally regarded as the gold standard of primary care databases – was the ideal setting for those studies.[1]

The strengths of the CPRD were discussed in detail in Section 2.4, and include its large size, high data quality and the fact that it is a probability sample[a] of the UK population. Indeed, during the recruitment period of this study, more than 500 GP surgeries in England, Northern Ireland, Scotland and Wales were registered with the CPRD, covering approximately 5% of the UK population. Therefore, inferences can be drawn from the results of this study to the UK population as a whole.

13.2.2 Case finding

In order to identify and quantify recording style bias between cases and controls it was paramount to identify cancer patients (the cases) and healthy patients (the controls) accurately.

As discussed in Section 3.2.2, more cases of urinary tract and pancreatic cancers were recorded on the National Cancer Data Repository than in the CPRD.[25] In total, this affected a relatively small number of cases in that study; namely, 20 urinary tract cancer cases (so, presumably an even smaller number of bladder cancer cases) and 23 pancreatic cancer cases that were recorded in the cancer registry but not in the CPRD.

[a] Probability sampling means that every person in the population has a known – and non-zero – probability of being included in the sample selected for study. It is valid to make inferences about population parameters from probability sample estimates.

A recent population-based cohort study validated diagnoses of colorectal, lung, gastroesophageal and urological cancers in England recorded in the CPRD by comparison with Cancer Registry data. The study period – 1 January 2001 to 31 December 2007 – coincided with my study's recruitment period.[132] Validation of CPRD diagnoses was reported in terms of their positive predictive value (PPV) (namely the proportion of CPRD cancers confirmed in the Cancer Registry), their sensitivity (i.e. the proportion of cancers correctly identified in the CPRD) and specificity (i.e. the proportion of patients identified as cancer-free in the CPRD). While neither bladder nor pancreatic cancer was investigated specifically, the study reported that the PPV of a CPRD diagnosis was 92% for urinary tract cancers and 96% for gastroesophageal cancers. The sensitivity was 85% for urological and 92% for gastroesophageal cancers. The specificity was also high – at 99% for both cancer sites – suggesting low numbers of false-positives within the CPRD. Urological and gastroesophageal cancer diagnoses were recorded a median of 4 days (interquartile range −22 to +13 days) and 9 (−29 to +7) days later, respectively, in the CPRD than in the Cancer Registry.

The authors noted study limitations, such as use of a convenience sample[a] and restriction to CPRD practices in England with linkage to the Cancer Registry. Therefore, inferences about the CPRD overall should be made with caution.[132]

[a] Convenience samples are non-probability samples that are drawn from a population purely because they are close to hand and readily available.

Overall, the findings of both studies are reassuring that cancer diagnoses recorded in the CPRD can be viewed as valid and accurate approximately 90% of the time.[25,132]

13.2.3 Symptom identification

In order to identify recording style bias between cases and controls accurately, it was important to identify all coded and text records of each of the symptoms studied. My study has several strengths and some limitations in this regard, which are now discussed.

13.2.3.1 Symptom thesauri

An undoubtable strength of the study was the comprehensive nature of the symptom thesauri compiled to identify the frequency of patient attendance, where this had been recorded.

There were two main indications that the symptom thesauri were sufficiently comprehensive to identify all coded records of attendance for haematuria, jaundice and abdominal pain.

First, it appears that, despite having the choice of a variety of codes for each symptom within the Read code system, GPs consistently used a relatively small selection (see Section 9.1.3). The rare or ‘doubtful’ codes that were retained in the thesaurus were used very sparsely, or never at all.

Secondly, identifying the codes that were paired[a] with text-only records of attendance for haematuria, jaundice or abdominal pain incidentally afforded the opportunity to examine the scope of the symptom thesauri to identify *all* coded attendances for those symptoms. None of the paired codes were identified as having been overlooked for inclusion in the symptom thesauri. This was reassuring, particularly for abdominal pain where we had decided against including codes for abdominal discomfort and abdominal tenderness (see Section 6.8.1).

Strictly, identifying these paired codes does not wholly exclude the possibility that the thesauri had still omitted some codes that GPs had used to document the symptoms. However, it would be on the very unlikely condition that the GP had never commented further on the symptom in the paired text box.

It is, of course, possible that GPs omitted to record symptoms reported by patients and that patients themselves did not report symptoms they were experiencing. This study was limited in that it could not assess the extent of these unrecorded or unvoiced symptoms.

Despite the above-mentioned caveat, it is reassuring that the methods used to compile the symptom thesauri – adopted in all CAPER studies (Section 4.2) – were thorough. My study confirms that most, if not all, coded records of

[a] *Reminder*: In ViSion, the main IT program used by GP practices contributing to the CPRD, a text record can only be made once a code has been selected – hence, all text records must have a 'paired' coded entry.

attendance for symptoms that were both reported and recorded were extracted consistently and reliably in CAPER studies.[40,41,42,43,44,45,46]

13.2.3.2 Symptom identification from text records

There were some limitations to the methods used to identify text records of attendance for the symptoms examined in my study.

First, the CPRD methods used to identify extracts from the text fields will have missed any references to symptoms made using incorrect or US spelling. This is particularly relevant to haematuria, the American spelling of which is hematuria. Reassuringly, we found no American English spelling elsewhere in the text extracts (e.g. estrogen rather than oestrogen) and a very low typographical error rate of 0.64%. Even so, the frequency of attendance identified from text records is probably very slightly underestimated.

Secondly, only the search phrase plus three words either side was available for analysis. This sometimes made it difficult to decide whether the symptom was a current concern or not. For example, two practising GPs could not agree as to the meaning of 14 out of 100 randomly selected text extracts, and agreed that a further two extracts were uninterpretable. I decided to err on the side of caution and categorise patients as not having a symptom if there was any uncertainty. Therefore, the frequency of attendance from text records is likely to be underestimated.

Thirdly, the study did not examine whether the use of text-only recording changed over the course of the recruitment period, i.e. from 1 January 2000 to

31 December 2009, In hindsight, given the knowledge that the CPRD is no longer able to collect text data or allow access to historical text data, it would have been valuable to explore trends in text-only data recording over this 10-year period.

13.2.4 Lifestyle factors

Lifestyle factors, such as smoking and obesity, were not included as possible explanatory variables in the original studies. This was deliberate, as the studies were seeking to model the prodromal symptoms of bladder and pancreatic cancer rather than to identify risk factors for developing these diseases.

It is possible that a GP would consider cancer more likely in symptomatic patients with known risk factors than in symptomatic patients without these risk factors. If the summated perceived risk of cancer affected the recording style, then information about these risk factors would be helpful. That said, even if the smoking and obesity data were reliable and available, my case–control study design is not optimal for investigating the effect of these lifestyle factors on recording style bias. The direct investigation of factors that contribute to a GP's decision of whether to record a patient's symptom using a code or solely in the text would require a separate study, in which the outcome was the recording style rather than a cancer diagnosis.

I carried out some *post-hoc* analyses (see Section 13.3.9 for the discussion of these findings) to explore the potential of this approach for future studies. However, a moratorium on the collection of text data by the CPRD was

introduced in 2013, reducing the practicability of this as a future avenue of research.

In reality, the recording of body mass index tends to be opportunistic, occurring when patients attend for another reason or because it is directly relevant to the patient's current clinical condition. Furthermore, the body mass index changes over time; therefore, even if this information were recorded in the CPRD, it may not accurately reflect the patient's status at the time the GP was assessing the likelihood of cancer. Bhaskaran *et al*[133] assessed the completeness of body mass index recording in the CPRD in a random sample of 1 million patients aged 16 years or older. They reported that a recent value of body mass index (i.e. in the previous 3 years) had been recorded for only 34% of their sample in the years 2000–2004 and for 51% in the period 2005–2011, indicating the poor availability of this data type in the CPRD.

An incentive to record smoking status was introduced with the Quality and Outcomes Framework in 2004.[47] Booth *et al* (2013) conducted a validation study of smoking records in the CPRD between 2007 and 2011. They reported that, of their sample (*n* = 279,682) of patients from English practices, 226,568 (81%) had at least one record indicating their smoking status, suggesting that recording levels are high. In addition, there was good agreement between smoking prevalence estimates in the CPRD and in the Health Survey for England.[134] This suggests that it would be feasible to design a study to investigate the effect of smoking status on recording style bias, although it would have to be set prior to 2013 to match the period of text record availability.

Even so, such a design would require establishing each patient's past (rather than current) smoking habits to determine their level of exposure to this risk factor.

13.2.5 Possible sources of bias within the study

13.2.5.1 Recall bias

Recall bias is a common concern for case–control studies, which generally rely on participants' memories of past exposure to risk factors. While the analysis was conducted retrospectively, the consultations themselves were recorded prospectively; therefore, recall bias was eliminated from both this and the original studies.[4,5]

13.2.5.2 Reporting bias

Matching on each of sex, age and GP practice minimised the potential for reporting bias. However, one could argue that patients in the early presentation stage of cancer might be starting to become worried about their health and have a lower threshold for attending the GP than controls. In the original studies this was examined by looking for differences in attendance between cases and controls for fractures. This was based on the concept that the more frequently patients attended the surgery, the greater their opportunity to voice unrelated items in addition to those that prompted the consultation. Fractures were chosen because they were thought to be equally likely across the entire study population, given that incident rather than metastatic cancer was investigated.

Neither of the original studies reported evidence of differential reporting patterns between cases and controls for fractures.

The choice of fractures may be criticised, however, because they are unambiguous diagnoses requiring the patient to seek treatment and follow-up in fracture clinic. Arguably, a fracture is unrepresentative of 'low-risk but not no-risk' symptoms such as abdominal pain, the reporting of which may vary with a patient's individual tolerance, capacity for self-management or knowledge of the symptom's significance.

A study of the presentation of cancer in childhood, carried out in association with our group, used attendance patterns for head lice and acne rather than fractures to examine differential reporting between patient groups.[135] There was no evidence from either of these 'control' conditions to indicate reporting bias between the cases and controls.

13.2.5.3 Detection bias

In my study, the non-visible haematuria records are vulnerable to detection bias.[136] This is because non-visible haematuria is only detected by specific dipstick testing of the urine. As described in Table 4.6, specific indications for urine dipstick testing include screening for (and monitoring of) diabetes mellitus and the diagnosis of renal disease. Therefore, patients with diabetes mellitus or renal disease, such as glomerulonephritis, may be over-represented in the group identified as having non-visible haematuria. There is no reported association between diabetes mellitus or glomerulonephritis and bladder

cancer, suggesting that testing is likely to be equally spread between cases and controls. However, urine dipstick testing may also be prompted where the GP suspects urological disease (including bladder cancer), leading to preferential testing in cases over controls.

The detection bias probably leads to an underestimation of the total number of patients with non-visible haematuria. In addition, the effect is likely to be unequal between the cases and controls. This suggests that any detection bias present in my study may inflate the risk estimates for bladder cancer in patients with non-visible haematuria.

13.2.5.4 Selection bias

Another limitation of the study is that it did not identify whether the recording style bias was caused by the recording behaviour of a small number of GP practices. However, it was felt important to include data from all contributing practices in order to obtain a measure of bias in CPRD studies overall. This is because the CPRD does not provide quality measures about the quantity of data recorded in hidden text – arguably they should be encouraged to do so at the practice level. This is particularly important now that text data are no longer collected and will not be available in future studies.

13.2.6 Missing data

This was an observational study of electronic medical records made during routine clinical practice; therefore, information about particular symptoms of

interest was not specifically sought by GPs, generating much missing data. Missing data is an acknowledged limitation of the CPRD.[2]

In my study, the lack of either reporting or recording of a symptom was interpreted as no evidence that the symptom occurred. Altman and Bland caution against this approach,[137] presumably because it risks introducing misclassification, particularly for binary outcomes. Including text records goes some way to correcting for any such misclassification, although it cannot correct for errors that arise when patients either do not voice their symptoms, or GPs do not record them by any method.

Herrett *et al* acknowledge that the approach taken in my study is the only workable option when faced with missing Read codes for binary variables in CPRD studies.[2] An alternative is imputation, but this is only suitable for continuous variables. Even then, patterns of 'missingness' vary and are complex in the CPRD, making it difficult to impute the missing values of continuous variables accurately. Herrett *et al* cite the examples of body mass index, which is more likely to be recorded in patients with ill health, and of blood pressure, which is recorded more frequently in those with cardiovascular disease. Another alternative is to restrict analysis to patients with complete data; however, this would drastically reduce the power of the study and introduce a bias in favour of patients who have ill health and who attend frequently.[2]

13.3 Discussion of methodological findings

This section starts with a brief overview of the symptoms to frame the ensuing discussion of the quantity of information about the symptoms recorded in hidden text at the event level and the consequences of this at the patient level.

13.3.1 The symptoms – a brief overview of their clinical significance

13.3.1.1 Jaundice

Jaundice has a variety of causes in adults, including gallstones and life-threatening malignancies such as pancreatic cancer. Gallstones typically present in women over 40 years old, who are both overweight and have a history of several pregnancies[85]; therefore, they were unlikely to be a common cause of jaundice in the pancreatic cancer study.[a]

As jaundice may indicate the presence of a serious condition, it is generally viewed as a worrying – 'alarm' – symptom in adults, particularly in those free both of gallstones and alcoholic disease, because of the possibility of pancreatic cancer.[9,93]

Jaundice has no reported association with bladder cancer.

[a] Indeed, in my reanalysis of the pancreatic cancer dataset, I identified that only 138/20,094 (0.7%) of the patients had a history of gallstones, of whom 83 (60%) were women.

13.3.1.2 Visible haematuria

Visible haematuria is also an alarm symptom with a variety of causes, ranging from urinary tract infections and calculi to serious malignancies such as bladder cancer. Benign nephrological causes, for example calculi, generally present in patients younger than those recruited to my study, who were aged 40 years or over.[a] In contrast, urinary tract infection is common in women, owing to their short urethra, and older women are particularly susceptible because their oestrogen levels fall after the menopause.[85] This is exemplified by the fact that, in the bladder cancer study, of the patients who experienced a urinary tract infection in the analysis period, nearly half were women, even though they only made up about one-quarter of the overall study population.[b] Urinary tract infection was associated with bladder cancer strongly enough to be included in the final model of risk markers of the disease (see Table 10.9).

Visible haematuria is also viewed as a serious clinical symptom, particularly in patients who either present with other features of urological tract malignancies or in those confirmed as not having a nephrological cause, urinary tract infection or stones.[72] The updated NICE guidelines recommend referral for patients for

[a] In my reanalysis of the bladder cancer dataset, I identified that only 12/26,633 (0.05%) of the patients had an episode of calculi during the analysis period.

[b] Of the 1,540 patients who had a urinary tract infection in the analysis period in the bladder cancer study, 707 (46%) were women, which is high considering that women comprised only 7,618/26,633 (28.6%) of the study population.

suspected bladder cancer if they are aged 45 years and over and have unexplained visible haematuria without a urinary tract infection, or persistent visible haematuria that recurs after successful treatment of a urinary tract infection.[9]

There are similarly no reported associations between visible haematuria and pancreatic cancer.

13.3.1.3 Abdominal pain

In contrast with these alarm symptoms, abdominal pain is a 'low-risk but not no-risk' symptom associated with both the cancers studied. In isolation, abdominal pain does not merit investigation for either bladder or pancreatic cancer; however, the updated NICE guidelines recommend that patients should be referred for possible pancreatic cancer if they present with abdominal pain in association with weight loss.[9]

Abdominal pain was the only symptom studied that was common to both cancer sites. Usage of text-only recording for this 'low-risk but not no-risk' symptom first appeared to be inconsistent between the cancers. Using the bespoke symptom thesauri from the original studies, text-only recording was used for a higher proportion of attendances in the bladder (1,106/2,795 = 40%) than in the pancreatic (2,215/7,191= 31%) cancer dataset. 'Abdominal pain', however, was defined far more broadly in the symptom thesaurus developed for the original pancreatic study compared with that created for the bladder cancer one, as it included codes for dyspepsia and indigestion. In contrast, identical search

criteria were used to extract the *text* records of abdominal pain from both cancer sites. As a consequence, text-only recording will have accounted for a smaller proportion of total abdominal pain records in the pancreatic, compared with the bladder cancer, study. Subsequent analysis using a generic symptom thesaurus for both cancers revealed that this 9% discrepancy in text-only recording between the cancer sites was an artefact, entirely accounted for by the differences in the bespoke symptom thesauri.

13.3.1.4 Non-visible haematuria

Non-visible haematuria is mostly associated with benign conditions such as glomerulonephritis, which generally present in patients younger than the participants recruited to my studies.[77]

In terms of cancer diagnosis, non-visible haematuria is viewed as another 'low-risk but not no-risk' symptom. Nevertheless, the NICE guidelines were recently updated[a] to include a new recommendation that patients who are aged ≥60 years and who have unexplained non-visible haematuria and either dysuria or a raised white cell count on a blood test are referred for suspected bladder cancer.[9]

[a] Using evidence in a paper published from this thesis.

13.3.2 Quantity of information recorded in hidden text at the event level

At the event level, differing (and large) numbers of attendances for each of the four symptoms – haematuria (visible and non-visible), jaundice and abdominal pain – were recorded solely in hidden text (see Figure 9.1 and Figure 9.2). Therefore, the original studies could not have been aware of the reported and recorded occurrences of these symptoms.[4,5]

13.3.2.1 Clinical significance of the symptom

The amount of text-only recording was independent of the symptom's clinical significance in terms of malignancy. Take, for example, the recording patterns for non-visible haematuria and jaundice – polar opposites in this regard.[72,77,93] Text-only recording was used frequently for both these symptoms, documenting well over half of the attendances for non-visible haematuria (298/494 = 60%) and for approximately half of the attendances for jaundice (1,639/3,338 = 49%). In contrast, a far lower proportion of attendances for visible haematuria – another alarm symptom, albeit of lower risk than jaundice[72] – was recorded solely in the text (2,699/7,302 = 37%).

13.3.2.2 What are GPs coding in preference to the symptom itself?

13.3.2.2.1 The source of information

Analysis of the code paired[a] with the text-only recording suggested that, at the event level, documenting the source of symptom information (frequently 'Letter from Specialist') using a readily retrievable method (i.e. codes) is an important factor for GPs.

Good-practice guidelines – current during the study period – were drawn up to advise GP practices as they migrated over to paperless systems. These guidelines give some insight into why GPs frequently gave priority, in terms of retrievability, to the information source over the information itself. Specifically, Sections 4.13 and 4.14 in the guidance stipulate the requirement to ensure that contacts, encounters and interventions that take place outside the GP surgery are clearly recorded using location codes.[138] This information is undoubtedly important, as it presents a complete picture of the care received by a patient.

13.3.2.2.2 The context of the consultation

At the event level, text-only records for all symptoms were also frequently paired with codes describing the context of the consultation. This may be

[a] *Reminder*: In ViSion, the main IT program used by GP practices contributing to the CPRD, a text record can only be made once a code has been selected – hence, all text records must have a 'paired' coded entry.

because GPs want to initiate the medical record at the outset of the consultation (which, in ViSion, requires selection of a code) and are unable to select an appropriate clinical code until they have taken the history.

In some cases, it may be that documenting the context of the consultation in a clear and easily retrievable way reflects defensive medical record-keeping practices, with GPs ensuring that they have defined the limits of their clinical interaction with the patient. The frequent pairing of 'Telephone encounter', for example, with text-only records of all symptoms likely reflects the importance for GPs of logging that fact that their ability to carry out a full and thorough examination was diminished.

13.3.2.3 The increasing importance of 'retrievability'

When medical records were entirely paper-based,[a] there was no choice but to retrieve information manually. The issue of 'retrievability' has grown in importance with the increasing computerisation of medical records. This is not only because computer searches can be used to automate, quickly and at reduced cost, the retrieval of anything recorded using a code. It is also because, when they are creating the electronic medical record, GPs now have to consider what information needs to be easily retrievable (i.e. coded), by whom and to what end.

[a] In primary care, paper records were stored in 'Lloyd George Envelopes'.

13.3.2.3.1 Updated best practice recommendations for record-keeping

The recommendations for good record-keeping that were in place during the study period[138] were updated in 2011, and have elevated the importance of 'retrievability'.[139] The updated guidelines state that 'where information can be adequately recorded using codes and structured data entry it is generally better to do so, but where this is not possible free-text clinical narrative can be used instead of or to clarify structured data entry.'[139] In this context, 'clear documentation' means 'coding', as this is the only recording method that ensures the information is readily retrievable in a computerised search. It is, of course, moot whether GPs ever read, let alone adhere to, such lengthy documents.

The updated recommendations reiterate the importance of clearly documenting the source of all clinical information, to capture that which comes from outside the practice. Acknowledging the impracticality of processing what may be large quantities of information, the guidelines recommend recording key data using a code.[139]

Finally, the updated guidelines place increased emphasis on the importance of documenting the context of all encounters between GPs and their patients, both within and outside the GP surgery. The guidelines state: 'when information is recorded that is likely to be shared with others working in a different setting, particular care needs to be taken to ensure that important context is made as explicit and unambiguous as possible'.[139] This assists care where patients are

often seen by a number of different healthcare professionals or GPs within a practice.

13.3.2.3.2 Medico-legal factors

An increasingly litigious atmosphere is also driving GPs to ensure that medico-legally important information is readily retrievable (i.e. coded) from the electronic medical record.

The numbers of complaints and medico-legal claims against GPs is increasing year on year.[140] In addition, the Criminal Justice and Courts Act 2015, which came into force in April 2015, extended the remit of the criminal offence of 'ill-treatment or wilful neglect' from children and people who lack capacity to cover competent adults.[141] An important aspect of this extension is that a patient need only complain about aspects of the clinical care, or about a near miss, for the event to be considered a possible 'crime'. While the Medical Defence Union anticipates that the offence of ill-treatment or wilful neglect will be prosecuted rarely, they predict an increase in criminal investigations of GPs.

In addition, new legislation was introduced in 2014, placing a statutory requirement on all doctors to have relevant insurance or professional indemnity to cover their medical practice. The majority of doctors had already made such provision, but its profile was raised by the legislation's granting of power to the General Medical Council to remove a doctor's licence to practise if such cover is not in place.[142] This adds further impetus to the legalistic view of the medical

consultation, and its subsequent recording. The maxim ‘if it isn’t written down, it didn’t happen’ is at risk of mutating into ‘if it isn’t coded, it didn’t happen.’

13.3.2.4 Summary and implications

The growing pressures to prioritise the documentation of ‘governance-type’ events in a readily retrievable way raise a serious concern that increasing quantities of important clinical information about patients will be unavailable in the future to both clinical audit and researchers because its recording is ‘relegated’ to the hidden text.

The above discussion has centred on the recording of symptoms at the event level, and how text-only recording may lead to underestimation of the number of episodes of a symptom. Re-attendance at the GP practice offers patients another opportunity to voice their symptoms and for the GP to record them in a retrievable manner using a code. This brings me to the discussion of my analysis at the patient level.

13.3.3 Quantity of information recorded in hidden text at the patient level

My study shows that GPs frequently and consistently recorded attendances for haematuria, jaundice or abdominal pain in the hidden text. For some patients, this occurred to such an extent that their entire record of attendance for that symptom was concealed from the original studies, which had analysed just the

coded records.[4,5] This led to their underestimating the numbers of patients with a history of these symptoms.

13.3.3.1 The importance of a symptom's clinical significance

Moving on from event-level to patient-level analysis,[a] a pattern appeared in the bladder cancer study suggesting the importance of a symptom's clinical significance. This contrasts with event-level data, where text-only recording appeared to be independent of clinical significance, driven rather by the source of symptom information and the context of the consultation.

The results from the bladder cancer study suggested that underestimation of the number of patients with a history of a symptom was inversely proportional to the risk of bladder cancer associated with that symptom. Compare, for example, visible and non-visible haematuria – conventionally perceived as contrasting symptoms in terms of bladder cancer risk. Within bladder cancer, for nearly 60% of the attendees for the low-risk, non-visible form of haematuria, all their records of attendance for the feature were concealed in the text,[77] compared with only 20% of all attendees for the 'alarming' visible form.[72] The 35% of attendees for 'low-risk but not no-risk' abdominal pain identified solely by text records lay between these values (see Figure 9.4).

[a] *Reminder*: Each patient was categorised by the overall style used to record their attendances for the symptom. 'Coded' was assigned if *any* record of a symptom was in coded form; conversely, 'text-only' was designated only when *all* instances were noted solely in the text.

These findings are entirely consistent with the hypothesis that doctors preferentially record clinical features substantiating what they think is the correct diagnosis and, furthermore, that they are more likely to use codes than text for alarm symptoms (see Section 5). This was also explored in *post-hoc* analysis (see Section 12), the results of which are discussed below in Section 13.3.9.

In contrast, the results from the pancreatic cancer study were not in keeping with the above-mentioned hypothesis, and possible reasons for this are discussed below.

13.3.3.2 In the pancreatic cancer study

The pattern of recording in the pancreatic cancer study was not consistent with the hypothesis that doctors preferentially record clinical features substantiating what they think is the correct diagnosis, and that they are more likely to use codes than text for alarm symptoms.

Rather, records for 30% of patients attending for jaundice (a ‘red flag’ for pancreatic cancer) in the analysis period were concealed in the text, compared with records for 22% of patients attending for abdominal pain (see Figure 9.4).

There are several possible explanations which, collectively, may account for this discrepancy between cancer sites, as discussed below.

First, for 77 cases (77/3,635 = 2.1%), their only documented presentation with jaundice coincided with their diagnosis with pancreatic cancer. For these 77 patients, it was the diagnosis that was codified into the electronic medical

record – the jaundice was noted solely in the accompanying text field. Arguably this reflects the greater significance of the diagnosis compared with its symptomatic manifestation. In contrast, a smaller proportion of cases in the bladder cancer study (43/4,915= 0.9%) had a single episode of visible haematuria that not only coincided with their date of diagnosis but was also noted solely in the text. Conceivably the difference between the cancer sites here reflects the fact that jaundice is generally a late feature of pancreatic cancer.[143]

Secondly, the symptom thesaurus for abdominal pain was broader in the pancreatic than in the bladder cancer study, including as it did codes for dyspepsia and indigestion. In contrast, identical search terms were used across the two studies to extract text records of the symptom. Therefore, by default, the proportion of attendees for 'abdominal pain' in the pancreatic cancer study identified by text-only records will be artefactually low. Indeed, using the generic symptom thesaurus for abdominal pain increased the proportion of attendees whose entire record was documented in the text in the pancreatic cancer study from 22% to 25%.

Finally, re-attendance for abdominal pain was far more common in the pancreatic than in the bladder cancer study. This is likely to reduce the proportion of attendees whose entire history of the symptom was documented solely in the text, simply because re-attendance for a symptom increases the number of opportunities for GPs to codify the symptom into the electronic medical record. It may, however, also reflect the fact that GPs are recognising

the increased clinical significance of the abdominal pain with re-attendance, and deliberately ensuring its documentation in a retrievable manner using a code.

This brings us to the importance of the context of presentation.

13.3.3.3 Context of presentation

The ‘context of presentation’ includes considerations such as the likelihood that the patient’s symptom(s) manifest serious disease, the constellation of other signs and symptoms presented by patients during a consultation, as well as patient characteristics such as their age.

If the context of presentation played no role in determining how symptoms were recorded at the patient level, one would predict that their recording patterns would be identical across the bladder and pancreatic cancer studies.

13.3.3.3.1 Age

The age profiles of the participants in the two cancer studies were very similar: both studies shared the same age criteria for recruitment, i.e. cases had to be at least 40 years old, and the controls were matched with the cases on age. Furthermore, the median age of diagnosis was 74 (95% CI: 66–80) years for bladder cancer cases and 73 (65–80) years for pancreatic cancer cases. Therefore, any difference in context of presentation between the studies is unlikely to be due to variation in the age profiles of their respective participants.

13.3.3.3.2 Alarm symptoms

One-fifth of attendees (696/3,483 = 20%) for visible haematuria had their entire history of attendance for the symptom during the analysis period documented solely in the text in the bladder cancer study. This fraction more than doubled *for the same symptom* in the pancreatic cancer study (142/341= 42%). A similar pattern was observed for jaundice in the pancreatic and bladder cancer studies, respectively (see Figure 9.13).

The results suggest that, once easily explicable causes have been excluded, the GPs are making strong clinical judgements that the possibility of pancreatic or bladder cancer has increased in patients presenting with jaundice or visible haematuria, respectively. Once they have recognised the seriousness of this possibility, GPs may ensure that this is documented in a readily retrievable manner in the patient's notes, hence the increased preference for coding.[139] This discussion is expanded below, where patient factors are explored at the level of cases and controls (see Section 13.3.4).

13.3.3.3.3 'Low-risk but not no-risk' abdominal pain

The trend for abdominal pain recording was similar in direction – but not magnitude – to that described above for visible haematuria and jaundice. Of the two cancer sites, the bladder cancer study had the greatest proportion of

patients with a history of abdominal pain concealed in the hidden text, reflecting the comparatively weaker association between them (see Figure 9.13).[a]

In contrast with the alarm symptoms, however, this effect was small in size (4% difference between cancer sites). Consequently, despite its reaching statistical significance, it was of questionable clinical relevance. One possible explanation for the reduced effect seen for abdominal pain, compared with the alarm symptoms, relates to the fact that abdominal pain is such a non-specific complaint, particularly in the primary care setting.[85]

13.3.4 Are studies of coded data vulnerable to bias arising from the differential use of text and codes between comparison groups?

Unbiased estimates of clinical features in diseased and healthy groups require that GPs are equally likely to record information reported by the two comparison groups. Not only that, the accessibility of information within the electronic medical record must be the same for both groups. As discussed in detail above,

[a] The same, generic, abdominal symptom library was used to compare how abdominal pain was recorded at the patient level between the bladder and pancreatic cancer studies. This is because re-using the bespoke symptom thesauri created for the bladder and pancreas studies generated a spurious difference in usage of text-only recording of abdominal pain at the event level between the cancer sites.

in the CPRD, only the coded records are readily and routinely available to researchers, the text being generally withheld (see Section 2.2). I have discussed the impact of this at the event and patient level, and now move on to discuss patient factors in greater depth by examining differences between comparison groups, i.e. cases and controls.

A common assumption made by studies using electronic primary care data is that GPs have no systematic tendency to alter their recording style according to a patient's later-designated case/control status[4,5,40,41,42,43,44] – an assumption this study was specifically designed to test.

The results reinforce the earlier proposed hypothesis (see Section 5) that GPs are making strong clinical judgements (driven by clinical context) about the probable significance of symptoms in patients. GPs appear to preferentially code clinical features they consider significant to a diagnosis, while tending to use hidden text to record those that they think are not. Therefore, studies of coded data are likely to be vulnerable to bias arising from the differential use of hidden text and codes between comparison groups, depending on the nature of the symptom being investigated and the context of its presentation.

13.3.4.1 For alarm symptoms

As discussed above (see Section 13.3.3), the original studies of bladder[4] and pancreatic[5] cancer considerably underestimated the numbers of patients attending during the analysis period for the alarm symptoms of visible haematuria and jaundice. Here I discuss how this extended to their obtaining

biased estimates of the numbers of cases and controls attending the GP for an alarm symptom associated with that cancer. While the mechanism of bias was slightly different in each cancer dataset, the net result in both instances was the same: the original studies obtained more complete information about alarm symptoms in their cases than in their controls.

As discussed above, a greater proportion of attendees for visible haematuria in the bladder cancer study were identified by codes than by hidden text. This effect was more marked in the cases (with codes identifying 2,591/3,147 = 82% of attendees) than in the healthy controls (196/336 = 58%). Consequently, compared with the cases, a greater proportion of the visible haematuria records of control patients were lost in the hidden text.

An even greater bias towards preferentially identifying cases over controls with a history of jaundice was observed in the pancreatic cancer study. The increase in bias was driven by the greater propensity of GPs to ‘relegate’ jaundice to the hidden text in the control patients, to the extent that coding only identified one-third (10/31= 32%) of these patients, compared with 1,102/1,565 (70%) of the cases. This was a particularly revealing finding, given that the overall patient-level analysis in the pancreatic cancer dataset had initially appeared inconsistent with the hypothesis that GPs are preferentially coding symptoms that substantiate their working diagnosis (see Figure 9.4 and Section 13.3.3.2).

These results strongly suggest that GPs were correctly identifying the clinical significance of the alarm symptom (visible haematuria or jaundice) in the cases – all diagnosed with an associated cancer within the year – and preferentially

coding this into the medical record. Consequently, these recording style decisions introduced a bias that exacerbated the difference in burden of the alarm symptom between the comparison groups.

The repercussions of these findings for outcome measures such as the positive predictive value and odds ratio are discussed below (see Section 13.4).

13.3.4.2 For ‘low-risk but not no-risk’ symptoms

The original studies also underestimated the numbers of patients who had attended for the ‘low-risk but not no-risk’ symptom of abdominal pain, because, for large numbers of patients, the entire record of attendance for the symptom was made solely in the text. That said, there was no evidence that omitting text-only records from analysis led to biased estimates of the numbers of cases and controls with a history of abdominal pain. A similar pattern was observed for non-visible haematuria in the bladder cancer study.

Abdominal pain is common to many diseases, and is not exclusively associated with either bladder or pancreatic cancer. Therefore, this lack of bias between cases and controls does not refute the hypothesis that GPs are correctly identifying the clinical significance of the symptom and opting to codify it in the electronic medical records.

13.3.5 Does the recording style vary with the clinical context of presentation of a symptom?

As discussed in Section 5, GPs draw on their knowledge and experience to formulate a working diagnosis that best fits the patient's presentation. This led to the hypothesis that doctors preferentially record any clinical features substantiating the working diagnosis, as opposed to everything reported by the patient. Analysis of recording methods allowed me to take this further and deduce that doctors preferentially 'code' those clinical features that substantiate their working diagnosis, rather than record them in the hidden text.

The results at the patient level (Section 13.3.3.3) support the hypothesis that such recording behaviour leads to bias in the capture of information about symptoms based on their perceived likelihood of presentation in a particular context (Section 5). The bias caused the original studies to obtain more complete information about the alarm symptom in their cases than in their controls.

More detailed analysis at the level of cases and controls further supports and develops the conclusions drawn from results at the general patient level. Not only are GPs accurately identifying when alarm symptoms are manifestations of their associated cancer, they are also recognising when the alarm symptoms are symptomatic of a pathology other than cancer.

Take the recording patterns of visible haematuria as an example: the proportion of controls whose attendance(s) for the symptom was recorded using a code

was similar in the bladder (196/336 = 58%) and pancreatic (147/238 = 62%) cancer datasets. In contrast, the proportion of cases whose attendance(s) for the symptom was recorded using a code was far greater in the bladder (2,591/3,147= 82%) than in the pancreatic (52/103= 50%) cancer dataset. This loss of preference for coding visible haematuria in the pancreatic cancer cases actually reversed the recording style bias to the extent that more complete information was obtained about the alarm symptom in the controls than in the cases. A similar reversal of bias was observed for jaundice in the bladder cancer dataset.

13.3.5.1 Summary

The results at the patient level strongly support the hypothesis that GPs consider the constellation of signs and symptoms presented by patients to formulate a working diagnosis. Subsequently, they preferentially record in code any clinical features that substantiate this working diagnosis. This introduces a recording style bias into clinical research studies whose analysis is restricted to coded records. Of course, GPs also consider the patient's test results, but these are automatically coded into the patient's record such that these data are not vulnerable to any recording style bias.

The next section proposes a framework of factors that are likely to influence GPs when deciding whether to record symptoms using a code or in the text. It is suggested that this framework may be used by researchers to anticipate the size and direction of any recording style bias in studies of CPRD coded data or similar datasets.

13.3.6 Recording method choices reflect the balance between several pressures

The recording method decisions made by GPs can be considered in light of the strategies they take to arrive at a diagnosis.[108] The following discussion will focus on gut feelings,[144,145,146,147] heuristics,[110] hypothetico-deductive reasoning versus pattern recognition[109,148] and Kahneman's two-system approach to how humans form judgements and make decisions.[149,150] [149]

13.3.6.1 Gut feelings

Gut feeling is defined as a feeling of unease that something is wrong in the absence of specific indicators. In a small qualitative study, Stolper suggested that GPs experience two kinds of gut feelings during consultations. The first is a 'sense of alarm', even about patients without specific signs or symptoms indicative of serious disease, while the second is a 'sense of reassurance' when GP feels secure about the prognosis and therapy.[145]

Van Den Bruel et al.[151] devised a triage instrument for assessing serious infections in children and reported that the gut feeling that 'something is wrong' proved to be the best clinical indicator that a child was really sick, although it was unclear upon which signs and symptoms the physicians were basing their decisions.

The above studies suggest that even a 'sense' of alarm may encourage GPs towards investigation or referral, in effect lowering their threshold for taking

action. However, it is difficult to anticipate how this would influence what and how GPs record in the medical record, given that their gut feeling is not dependent on the presence of specific signs or symptoms.

13.3.6.2 Heuristics

Heuristics are simple decision strategies that focus on a small number of relevant signs or symptoms that are predictive of a diagnosis, ignoring some of the information available. Medical decision-making heuristics commonly follow a 'fast-and-frugal' model, whereby doctors' decisions are guided by the answers to a number of mostly subliminal, sequential, yes/no questions. [110]

Heuristics are based on three rules:

1. *The search rule*: this specifies what information is sought (i.e. a clinical predictor) and in what order of clinical priority.
2. *The stop rule*: this determines when the clinician can stop seeking new information (i.e. when sufficient clinical predictors have been obtained).
3. *The decision rule*: this determines the decision to be made, according to the value of the clinical predictor(s).

It is conceivable that these rules influence a GP's choice of what to code into the patient's medical record, with preference given to high-priority clinical indicators. Indeed, the results of my thesis support this hypothesis: visible haematuria and jaundice are both high-priority clinical indicators – the former for bladder cancer, the latter for pancreatic cancer – while abdominal pain is a 'low-risk but not no-risk' symptom of both cancers.

While heuristics may be useful to doctors working in time-pressured environments, they may also lead to errors in diagnosis or management. [152] For this reason, it is important that any heuristic frameworks should be validated and only used if they are shown to have a high sensitivity; however, this may come at the expense of a low positive predictive value, which has the potential to lead to over-investigation.[144]

Heuristic frameworks have many attractions; for example, their degree of accuracy can be validated, and they are transparent and easy to apply in appropriate settings. Furthermore, they may lead to cost savings, if shown to be as, or more, accurate than complex alternatives. [110] However, as acknowledged in the recently updated NICE guidelines for suspected cancer,[9] they are too simplistic to use in isolation,[108] and should be used in conjunction with a range of clinical decision-making skills, including hypothetico-deductive reasoning and pattern recognition, which are discussed next.

13.3.6.3 Hypothetico-deductive reasoning and pattern recognition

When assessing patients, a hypothetico-deductive reasoning strategy may be adopted by GPs, in which they decide upon a number of possible diagnoses early on in a consultation and then test them through focused examination and history-taking. An alternative is pattern recognition, in which GPs readily recognise a patient's presentation and make a 'spot' diagnosis. Under these circumstances, GPs may consider it unnecessary to 'test their hypothesis', reserving this strategy only for more complex or atypical presentations.[108,148] When assessing the probability of a diagnosis using either hypothetico-

deductive reasoning or pattern recognition, GPs's decisions are vulnerable to a number of cognitive biases, which may lead to diagnostic errors, including[148,153]:

1. Availability bias: where GPs are more likely to consider diagnoses that are more easily retrieved from memory
2. Representativeness bias: where GPs place too much emphasis on the typical features of a disease and miss an atypical presentation
3. Confirmation bias: where GPs only seek information to confirm their hypothesis, rather than also testing whether it might be incorrect
4. Premature closure: where GPs come to a decision too soon, before eliciting enough information (this is also pertinent to heuristic models)
5. Base rate neglect: where GPs neglect the true rate of a diagnosis and focus on a possible, but unlikely, diagnosis.

These biases, particularly confirmation bias, may also influence GPs' decisions about what to record in a patient's medical record, and whether they record it using a code or in the text (see Section 6.5.1.2).

Hypothetico-deductive reasoning and pattern recognition build on tangible clinical information obtained when GPs take a history and perform a clinical examination. They fit well with the work of Kahneman, which is discussed below.

13.3.6.4 Kahneman’s two-system approach

In his book, *Thinking, Fast and Slow*, Daniel Kahneman uses a ‘metaphor of two agents’ – which he names System 1 and System 2 – to characterise mental

thinking broadly as either intuitive or deliberate. He proposes that thought processes in System 1 occur automatically and quickly, with little or no effort and no sense of voluntary control. System 1, he says, underlies many of our inherent skills, ranging from the ability to judge distance to the recognition of stereotypes.

In contrast, System 2 is used for complex mental effort and when checking facts or logic, such as when testing hypotheses. Indeed, tasks carried out using System 2 are disrupted when a person's attention is diverted to something else, at the expense of performance of the task in hand.

Systems 1 and 2 both function while we are awake, but it is System 1 that automatically determines our impressions, intuitions, intentions and feelings related to events or situations we encounter. Through an iterative process, System 2 reviews the interpretations made by System 1, and intuitions or impressions that are 'approved' by System 2 are reinforced to the extent that they become beliefs. Similarly, intentions and feelings that are substantiated by System 2 turn into impulses and eventually become voluntary actions.

In Kahneman's model, Systems 1 and 2 are coordinated efficiently to minimise effort while at the same time optimising performance. Indeed, most of the time System 2 is not very active – it is only alerted when something happens that 'violates the model of the world that System 1 maintains'. In other words, it is mobilised when System 1 does not have a ready explanation for what is going on around us.[150]

Kahneman's theory of System 1 and System 2 fits well with the above-discussed concepts of clinical reasoning, with System 1 equating to 'pattern recognition' and System 2 to 'hypothetico-deductive reasoning'.[148,149,152]

During consultations, GP's decision-making is likely to be dominated by pattern recognition (System 1), in part because short appointments do not allow for the slow thought processes of System 2. Indeed, through repeated learning and development of expertise and specialist knowledge, doctors have gained, through System 1, the ability to store patterns of association in memory that are then accessed and recalled without intention or effort. Hypothetico-deductive reasoning (System 2) is reserved for complex patients, whose symptoms and history do not fit into an immediately recognisable pattern.[148,154] Some commentators, for example Marcum,[154] have suggested that there is a third process involved – termed 'metacognition' – in which doctors reflect on their thinking and decisions, particularly when there is a conflict between Systems 1 and 2. I suggest that metacognition plays a minor role in the short and pressured consultation itself; rather, it is more likely to occur later when GPs are pondering over difficult cases, or sharing them with colleagues.

Knowledge of the many cognitive influences described above on GPs during the patient consultation will help researchers to predict the presence and direction of any recording style bias at the patient level. My results strongly suggested the importance, at the patient level, of a symptom's clinical significance in relation to the disease under study.

GPs are mindful of alarm symptoms because pattern recognition alerts them of their association with serious disease (‘Framing & norms’ in Figure 13.1). Indeed, referral for alarm symptoms alone is encouraged by heuristics such as NICE guidance for cancer (‘Dogma’ in Figure 13.1).[9] Independently of national guidelines, GPs are likely to feel uneasy (‘Affect heuristic’) about patients who present with alarm symptoms, and worry about the consequences – for them and their patient – of their missing or delaying a serious diagnosis (‘Loss aversion’). Repeat attendance for a symptom, particularly if an innocent explanation has been excluded, will exacerbate these feelings.

A symptom that is considered to be ‘low-risk but not no-risk’ will not be associated with such strong ‘framing & norms’ effects and so will not engender the same degree of uneasiness and loss aversion as an alarm symptom.

All of the above suggest that alarm symptoms, rather than ‘low-risk but not no-risk’ symptoms, will provoke GPs to take action, diverting doctors away from a watch-and-wait approach. Procedural requirements associated with that action, such as form-filling, may encourage GPs to record the headline symptom or administrative procedure using a code. Once this has been done, everything else may be recorded quickly in hidden text, including symptoms that are irrelevant to, or less critical for, the working diagnosis.

There are a number of external factors that may also influence what GPs decide to record in the medical record. For example, the Quality and Outcomes Framework and IT factors.

13.3.6.5 Quality and Outcomes Framework

GPs are incentivised to code anything covered by The Quality and Outcomes Framework (QoF).[47] The QoF was introduced as part of the new GP contract in April 2004. Its purpose was to improve the quality of general practice by rewarding GPs financially for implementing 'good practice'. While participation is voluntary, it is widely adopted as a means of securing income. Much of what it entails is classic System 1 thinking, which translates into a recording style that must enable the ready retrieval of that clinical information, i.e. coding rather than text. Evidence of compliance with QoF requires GPs to code clinical events that fall within named domains, of which there were four in 2004: clinical, organisational, patient experience and additional services. The symptoms I studied – haematuria, jaundice and abdominal pain – were not included in any of these domains and so their coding will not have been encouraged specifically by QoF.[155]

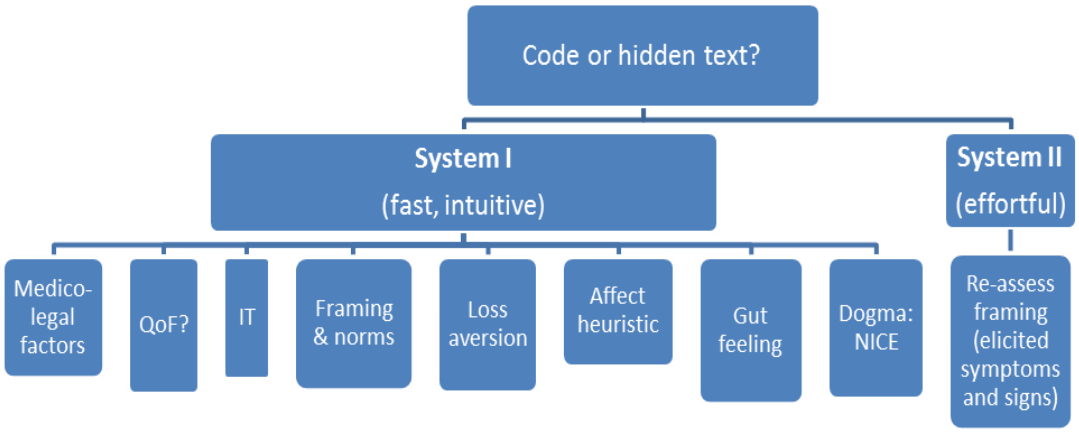
I suggest that researchers using CPRD datasets should first consider whether the symptoms they are studying were covered by QoF at the time of data collection. If they were, the researcher can be more confident that the majority of symptom occurrences will be coded consistently for all patients. Therefore, their estimates of both the frequency of symptoms and the number of patients affected by them will be good. Indeed, studies spanning the introduction of QoF may see a change in data recording quality, as suggested by a paper on ovarian cancer.[156]

For symptoms that are not covered by QoF, researchers should acknowledge the strong possibility that they will underestimate symptom frequency if they restrict their analysis to coded records. The reason for this is likely to be related to IT factors, which are discussed below.

13.3.6.6 IT factors

For anything outside QoF, time pressures encourage GPs to select the simplest method of recording it in the patient's notes (IT factors in Figure 13.1 below). This will vary with the GP's familiarity with the software, which could not be elucidated in this study. It will also vary with the utility of the software itself, which in most CPRD practices is ViSion. ViSion requires GPs to enter a code before the text box becomes available; therefore, at least one code must be selected for a consultation to be recorded. The event-level analysis suggested that researchers should not assume that the importance of a symptom will outweigh the importance of coding either the source of the information or the context of the consultation. Indeed, medico-legal factors will encourage GPs to document not only the symptom, but also its source and the context of the consultation, in a readily retrievable manner (i.e. as a code) at the event level.[139]

Figure 13.1 Recording style decisions: System 1 and System 2 factors



13.3.6.7 System 2 factors

Diagnoses rely not on a single symptom, but usually on a constellation. This is particularly so in cancer. I suggest that System 2 steps in to help GPs spot when an alarm symptom is a 'red herring' rather than a 'red flag'. This may be because something else in the patient's presentation contradicts the 'typical' diagnosis in patients presenting with that alarm symptom. Alternatively, it may simply be that the other signs and symptoms do not wholly fit with the recognised pattern of that disease. It may also be because the GP has a gut feeling of reassurance that the patient does not have serious disease,[146]

although electronic medical recording systems do not offer an obvious mechanism for recording this. My results suggest that, in this scenario, recording style bias is reversed in favour of obtaining information about the symptom in the controls rather than the cases.

This underlines the importance for researchers to relate the clinical significance of a symptom to the disease they are studying. It is difficult to envisage a scenario in which researchers studying a particular disease would want to investigate the frequency of alarm symptoms that are not known to be associated with that disease. In the unlikely event that they did, they should anticipate the presence of a recording style bias in favour of identifying symptom occurrence in the controls over the cases.

13.3.7 Added value of text data

This part of the discussion relates to research question 5: Do the text data provide additional value to coded data? The discussion starts with the value of text data in increasing the identification of symptom frequency, with particular reference to the 5% threshold that symptoms are required to exceed to be included in univariable analysis in some CAPER studies.

The discussion then continues by examining the ‘research quality’ of code- and text-based variables.

13.3.7.1 Threshold for inclusion in univariable analysis

As described in Section 6.12.1, only those symptoms, signs and test results that had occurred in at least 5% of the cases or controls were included in analysis. This somewhat arbitrary threshold was chosen as a pragmatic compromise. If it were any lower, the chances of identifying clinically meaningless associations were increased; any higher, and the chances of missing associations between important, yet infrequent, symptoms were raised.

In the original bladder cancer study extended by my PhD,[4] the frequency of non-visible haematuria (127/4,915 cases = 2.6%) failed the threshold for inclusion in analysis. Supplementing with text-only records for this feature revealed that its actual frequency was 312 in the 4,915 cases (6.4%) – above the 5% threshold. Therefore, inclusion of text-only records enabled the first study of the risk of bladder cancer in patients presenting with non-visible haematuria in primary care. The results from this thesis were published in *The British Journal of General Practice*[8] and used by the National Institute for Health and Care Excellence in their revision of the guidelines for recognition and referral for suspected cancer. A new recommendation was introduced to refer people for suspected bladder cancer if they are 60 years of age or more and have unexplained non-visible haematuria and either dysuria or a raised white blood cell count.[9] This is a direct demonstration of the added value of text-only records to research. These clinical findings are discussed in further detail below (see Section 13.5).

Arguably, one could also avoid missing such important features of cancer by lowering the threshold for inclusion to below 3%. Indeed, this has been done in other CAPER studies; for example, when determining the risk of breast cancer in symptomatic patients presenting to primary care.[42] However, this still leaves the risk estimates vulnerable to any recording style bias between cases and controls and is a less satisfactory solution.

13.3.7.2 The errors associated with code- and text-based variables

An undeniable disadvantage of the text-based variables compared with their coded counterparts is their relatively high degree of uncertainty. There were two main sources of this uncertainty. First, not all the text that had been written was accessible, just the search term and three words either side. The second source of uncertainty was the accuracy with which the text extracts were converted to binary variables, i.e. whether they denote a symptom's absence or presence. Several steps were taken both to minimise and quantify the degree of uncertainty associated with text-based variables.

13.3.7.3 Minimising the degree of uncertainty introduced by text variables

The classification of text extracts was carried out using a semi-automated process that ensured the consistent application of a set of classification 'rules' using an algorithm, and also allowed the flexibility of individual assessment where necessary. The arguments of the algorithm were underpinned by a solid

understanding both of syntax and the theory of negation (see Appendix 3: Algorithm construction).[112]

The main source of uncertainty in interpreting text extracts was whether the symptom was a current or historical concern for the patient. This was sometimes difficult to discern, but was important to establish for the purposes of the study, which was designed to estimate the risk of cancer in patients with a current symptom. The category 'symptom absent' was assigned if the meaning could not be ascertained with certainty, to minimise the number of false-positives in the dataset.

Increasing the number of words in the text extracts from three to five either side of the search term would improve the chances of being able to interpret the meaning accurately. At 5p per word, this would incur a considerable extra cost, for limited, if any, benefit.

13.3.7.4 Quantifying the degree of uncertainty

The performance of the semi-automated classification process was assessed by comparison of its output with that of a rigorously constructed gold standard, using a test set of 100 randomly selected text extracts.[113] Additional analyses were conducted to obtain maximum and minimum measures of sensitivity and specificity. Negation of the symptom in the text was rare, resulting in nearly all (90% in the randomly selected test set) the text extracts confirming its presence. This created an imbalanced contingency table, and rendered

interpretation of the sensitivity (which ranged from 96% to 100%) and of the specificity (100%) essentially meaningless.

A further indication of the degree of uncertainty in text-based variables was obtained during construction of the reference standard. This entailed two GPs independently rating a random sample of 100 text extracts, blind to whether they were from the medical records of cases or controls. The final level of agreement for all 100 extracts was 0.71 (95% CI 0.53–0.89), which, while not perfect, is interpreted as good by Altman[120] and as substantial by McGinn *et al.*[122] Furthermore, this value is likely to be an underestimate of the true level of agreement, because of the imbalance in the number of text extracts assigned to each category – far more were categorised as indicating the symptom's presence than its absence. For 16 of the 100 text extracts, the two independent raters could either not agree their meaning, or agreed that they were uninterpretable (these extracts were not included in the final reference standard). This in itself is an indication of the degree of uncertainty in the text-based variables.

13.3.7.5 The 'research quality' of code- and text-based variables

13.3.7.5.1 Modelling the presentation of cancer

Text- and code-based variables gave similar measures of association (i.e. odds ratio) between 'low-risk but not no-risk' symptoms and cancer. In contrast, the text-based variables gave smaller – but still significant – measures of association between recognised alarm symptoms and cancer (see Table 10.6

and Table 10.13). This fall in odds ratio was most likely to be an artefact related to the biased recording of alarm symptoms between cases and controls, wherein coded records obtain more complete information about the symptom history from cases than they do from controls. Therefore, the decision was made to combine the code- and text-based variables for each symptom into a single composite variable. This illustrates the value of text-based variables, as they 'correct' both for any recording style bias and underestimation of symptom frequency.

Further evidence of the added value of text-based variables was illustrated by the observation that their addition significantly improved the fit of the final models both of bladder (see Table 10.9) and pancreatic (see Table 10.16) cancer presentation.

13.3.7.5.2 Estimating the risk of cancer

Similarly, further value of text-based variables was apparent in their correction of the likelihood ratio and hence positive predictive value for those symptoms affected by recording style bias. This is discussed in greater detail below (see Section 13.4).

13.3.7.5.3 Diagnostic interval data

From the summary statistics and wide interquartile ranges, it appeared that adding text-only records had little impact on the estimate of diagnostic intervals (see Table 11.1). However, matched analysis of patients whose symptoms had

been recorded using both methods suggested that the addition of text-only records lengthened the diagnostic interval for all the symptoms in the cancer sites investigated (see Table 11.2). This was most marked for 'low-risk but not no-risk' abdominal pain in pancreatic cancer, where the median diagnostic interval increased by 17 days after addition of text-only records. It was least notable, and of questionable clinical significance, for the alarm symptom of jaundice in pancreatic cancer. Here, the median diagnostic interval increased by only 3 days.

Unmatched analysis was conducted using data from patients whose symptoms had been recorded only in the text or only as a code. The results were somewhat counterintuitive on first glance, given the results of the matched analysis. However, this probably simply reflects the fact that adding text-only records can either have no effect on the estimate of diagnostic interval, or lengthen it – it cannot shorten it. The unmatched analysis tells us whether diagnostic interval estimates from coded and text-only records are drawn from the same population.

The unmatched analysis suggested that coded records tended to be made *earlier* than text-only records for all symptoms, except visible haematuria in bladder cancer whose diagnostic interval estimates were the same regardless of recording style. As in the matched analysis, the difference in median values was smallest for jaundice in pancreatic cancer. At 4 days, the clinical significance of this change is questionable. The difference in median values

was greatest for abdominal pain, at 46.5 days in the bladder and 30 days in the pancreatic cancer datasets.

Overall, this suggests that, for alarm symptoms, the text-only records provide little added value in terms of estimating diagnostic interval. For 'low-risk but not no-risk' abdominal pain, however, the estimate of diagnostic interval is more accurate after the addition of text-only records.

To my knowledge, Hayward *et al*[157] is the only other comparable study. They reported that respiratory symptoms were recorded significantly earlier in the text than as a code. For example, the median diagnostic interval for the first coded symptom for breathlessness or wheeze was 44 (95% CI: 9–139) weeks, increasing to 149 (95% CI: 44–275) weeks after addition of text-only records. As far as I can tell from the paper, no statistical test was used to assess the strength of evidence of a real difference between these median values. Indeed, the large confidence intervals suggest to me that, while the point estimates appear different, there was in fact no difference between the two median values owing to the large variability inherent in the measurement. This finding certainly matches my observations of the raw data, as described above (see Table 11.1).

I conclude that, for the assessment of diagnostic intervals, text-based variables provide added value over coded records for 'low-risk but not no-risk' symptoms. In contrast, they provide little added value for the alarm symptoms.

13.3.8 Summary of methodological findings

My study shows that large numbers of individual symptom records are made solely in the hidden text. This suggests that researchers restricting their analysis of CPRD data to just the coded records need to acknowledge that their event-level reports of symptom frequency will be underestimated considerably. This study was unable to elucidate a framework that might be used to predict the extent and direction of this underestimation, as no association between use of hidden text and clinical significance was apparent.

Researchers should also acknowledge that this underestimation persists at the patient level, but may be compounded by a recording style bias between comparison groups. This is greatest for alarm symptoms strongly associated with the disease under investigation, where the bias preferentially obtains information from the cases compared with controls. For ‘low-risk but not no-risk’ symptoms that are not strongly associated with the disease under investigation, little or no recording style bias is anticipated.

The implications of these biases for clinical outcome measures are discussed below, but before that I would like to discuss the findings of my *post-hoc* analysis.

13.3.9 Discussion of *post-hoc* analysis

The *post-hoc* analysis was prompted by the finding that recording style bias for visible haematuria differed between men and women (see Section 9.2.2.1.1). It

also relates to the hypothesis that GPs tend to use codes for significant symptoms and text for less worrisome symptoms.

The *post-hoc* analysis suggested that patients were more likely to have their haematuria recorded solely in the text if they had a urinary tract infection. This is entirely consistent with the hypothesis that symptoms that commonly arise from benign pathology are more likely to be noted in the text fields than as a code. Since urinary tract infections are more common in women than in men, effect modification by gender of the association between urinary tract infection and text-only recording was anticipated, with GPs having a raised suspicion of a benign cause in women. The interaction term was not significant, however, which seemed counterintuitive. On reflection, this finding may indicate that GPs are only recording confirmed, rather than suspected, diagnoses of urinary tract infection.

Also consistent with the hypothesis was the observation that control patients – whose cause of visible haematuria was presumed to be benign – were more likely than bladder cancer cases to have their symptom recorded solely in the text. Here there was an effect modification by gender, with female control patients being three times more likely than male control patients to have their visible haematuria recorded solely in the text (see Table 12.6). Therefore, something about the women – but seemingly not the presence of a urinary tract infection – is increasing the likelihood that their records of attendance for visible haematuria are lost in the hidden text.

Possible reasons for this are now discussed.

13.3.9.1 GPs' perceptions of the risk of bladder cancer in women with visible haematuria

The finding of gender-based bias in use of text-only recording was specific to haematuria and was not observed for abdominal pain or jaundice. This observation may be explained by differences in GPs' perceptions of the risk of bladder cancer in men compared with women with visible haematuria.

For example, Lyratzopoulos *et al* reported evidence that GPs interpret the seriousness of haematuria differently in men than in women, suggesting that GPs are less likely to suspect a diagnosis of bladder cancer in women than in men. Furthermore, they found that women were more likely than men to have a delayed diagnosis of bladder cancer.[158] Their study was carried out using data on 740 patients with bladder cancer from the English National Audit of Cancer Diagnosis in Primary Care (2009–2010). One outcome measure used as evidence of gender inequality was the 'primary care interval'. This is the time interval between the first presentation to the GP with a symptom and the patient's first specialist referral for further investigation for cancer. There was only a small difference in the median primary care interval between men (4 days, *n* = 525) and women (6 days, *n* = 196). At the tail of the distributions, however, the difference was substantial. The primary care interval at the 75th centile was 15 days for men whereas it was 32.5 days for women, and at the 90th centile the difference was greater still, at 39 days for men and 103 days for women. In addition, women were more likely than men (odds ratio 3.29, 95%

CI: 2.06–5.25, p<0.001) to have visited the GP on three or more occasions before referral.[158]

A limitation of the latter study is that it did not report whether the 'excess' consultations by the women were for possible symptoms of cancer. Therefore, it is possible that some of the results could be accounted for by baseline differences in consultation rates between men and women. Unfortunately, Lyratzopoulos *et al* did not report a detailed age profile of the patients in their study, other than saying that 91% of the patients with bladder cancer were older than 55 years. As discussed in Section 6.4.1, aggregated data from QResearch® suggest that men aged 55–59 years consult the GP five times per year, compared with a figure of nearly seven for women.[101] In contrast, a THIN-based study reported that, in patients older than 58 years of age, there was little, if any, gender difference in consultation rates.[102] Gender-specific differences in baseline consultation rates are more likely to account for the 'excess' consultations by women observed in Lyratzopoulos *et al*'s study if they occurred over a long time period. If, however, they all happened shortly before the referral was made, one could be more confident that the 'excess' consultations in women truly reflected a gender bias in the GPs' interpretation of the seriousness of the patient's condition. Unfortunately, the intervals between consultations were not reported for this assessment to be made.

13.3.9.2 Misattributed vaginal bleeding

The women recruited to my study were all over 40 years old. Approximately one-tenth (639/6,266 = 10.2%) of the controls were between 40 and 59 years of

age, and likely to have been experiencing symptoms of the menopause during the study (see Table 7.3). One such symptom is abnormal and chaotic vaginal bleeding, which cannot always be distinguished from visible haematuria of urological origin.

It is not possible for GPs to verify the source of blood apparently observed in the urine (vaginal or urological) after the event, although the presence (or, equally, the absence) of other presenting features may raise their suspicion of misattribution of vaginal bleeding. Under these circumstances, it is possible, but I think unlikely, that the GP may record the patient's report of 'haematuria' in the text rather than as a code.

13.3.9.3 Summary

The results of the *post-hoc* analysis support the hypothesis that GPs tend to use codes to record visible haematuria when they suspect that they underlying pathology is serious, and text when they feel reassured of a benign cause. This effect was greater in women than in men, although a single reason for this gender bias could not be established.

The implications of the methodological findings in my main and *post-hoc* analysis for clinical outcome measures are now discussed.

13.4 Effect of recording style bias on clinical outcome measures

13.4.1 Likelihood ratio

This discussion is best started with a reminder of how the likelihood ratio was calculated (see Section 6.12.3.4 for more detail). In cancer diagnostics, the positive likelihood ratio summarises how many more times likely patients with cancer are to experience a symptom compared with healthy patients free of cancer. It is calculated using the following formula:

$$\frac{p(S+ \mid D+)}{p(S+ \mid D-)} \quad (1)$$

Where $p(S+|D+)$ is the probability that cases (i.e. patients who have a positive diagnosis of disease D, denoted as $D+$) have a symptom S ($S+$), and $p(S+|D-)$ is the probability that controls (i.e. patients who are not diagnosed with that cancer: $D-$) have the same feature, $S+$.

At the event level, use of hidden text in the electronic medical record conceals a considerable number of occurrences of all symptoms from researchers examining just the coded data. This persists with repeat attendance by individual patients, to the extent that the complete history of attendance for that symptom is also concealed from researchers at the patient level. Therefore, both $p(S+|D+)$ and $p(S+|D-)$ are likely to be underestimated for all symptoms.

13.4.1.1 For ‘low-risk but not no-risk’ symptoms

Similar proportions of cases and controls with abdominal pain and non-visible haematuria in the bladder cancer study were lost to research because their entire history for the feature was recorded solely in hidden text. In other words, the degree of underestimation of abdominal pain and non-visible haematuria was the same in cases as in controls. Therefore, the likelihood ratio estimate was unaffected by recording style bias; for example, the biased estimate for abdominal pain in the bladder cancer study was 2.02 (95% CI: 1.79–2.28) and the corrected estimate was 1.97 (1.79–2.17) (see Table 10.2). There was a small bias in estimates of abdominal pain in favour of identification of the symptom in the cases over the controls in the pancreatic cancer study. However, this was so small as to have a negligible impact on the likelihood ratio, the ‘biased’ estimate of which was 6.94 (95% CI: 6.46–7.45) compared with the corrected value of 6.59 (6.21–7.0) (see Table 10.11).

From both research and clinical perspectives, it is most reassuring to know that outcome measures for ‘low-risk but not no-risk’ symptoms are not likely to be affected by recording style bias.

13.4.1.2 For alarm symptoms in their associated cancer

When an alarm symptom is investigated in the context of a cancer with which it has a strong association, recording style bias leads to the preferential obtainment of information from the cases compared with controls. In terms of

equation (1) above, this means that $p(S+|D-)$ is underestimated to a greater extent than $p(S+|D+)$, *inflating* the likelihood ratio artefactually.

From a research perspective, studies should acknowledge that their likelihood ratios for *recognised* alarm symptoms of disease are likely to be overestimates. For example, in the bladder cancer study, the biased estimate of the likelihood ratio for visible haematuria was 58.41 (95% CI 50.69–67.32). Correcting for recording style bias by adding in the text-only records reduced this to 41.39 (37.14–46.11) (see Table 10.2). This also translates to an overestimation of the positive predictive value, as discussed below (see Section 13.4.2).

13.4.1.3 For alarm symptoms in unconnected cancers

It is unlikely that researchers will want to estimate the likelihood ratio of an alarm symptom in an unconnected cancer (see Section 13.3.6.7). If they were, researchers should be aware that recording style bias is likely to lead to the preferential obtainment of information from the controls compared with the cases. This means, in terms of equation (1), that $p(S+|D+)$ will be underestimated to a greater extent than $p(S+|D-)$, *reducing* the likelihood ratio artefactually.

13.4.2 Positive predictive value

As discussed in Section 6.12.3.5, the positive predictive value (PPV) was calculated from the posterior odds [PPV = posterior odds / (1 + posterior odds)], which was estimated using Bayes' Theorem:

Posterior odds = (prior odds × *f*) × positive likelihood ratio (2)

The adjustment factor, *f*, derives from the proportions of cases and of controls who consulted their GP for any reason during the period of the study. Therefore, it is independent of recording style bias related to individual symptoms. The value of prior odds is generally estimated from national incidence data, which again is independent of recording style bias. Therefore, the impact of recording style bias on PPV derives totally from its effect on the positive likelihood ratio (see Section 13.4.1).

The results strongly suggest that recording style bias is likely to inflate the PPV of alarm symptoms, but only for cancers with which they are associated. In contrast, recording style bias is unlikely to affect the PPV of 'low-risk but not no-risk' symptoms of cancer.

13.4.3 Odds ratio

In case–control studies, the strength of association between a symptom and cancer is estimated by the odds ratio: the greater the odds ratio, the greater the association between the symptom and cancer. Section 6.12.3.6 describes how the odds ratio is calculated using the formula:

$$\text{Odds ratio} = \frac{p(S+ \mid D+) \times p(S- \mid D-)}{p(S- \mid D+) \times p(S+ \mid D-)} \quad (3)$$

Where:

$p(S+|D+)$ = the probability that symptom S occurs ($S+$) in patients with cancer D (the cases, $D+$)

$p(S-|D-)$ = the probability that symptom S does not occur ($S-$) in patients who do not have cancer D (the controls, $D-$)

$p(S-|D+)$ = the probability that symptom S does not occur ($S-$) in the cases ($D+$)

$p(S+|D-)$ = the probability that symptom S occurs ($S+$) in controls ($D-$)

This was an observational study based on electronic medical records made during every-day clinical practice, i.e. information about the symptoms of interest was not sought specifically. Symptom absence was assumed if there was no evidence to the contrary (see Section 13.2.6 for discussion of the handling of missing data). Therefore, values for $S-$ were derived entirely from the estimates of $S+$ and the number of cases and controls.

From equation (3), it can be seen that a recording style bias favouring obtainment of information about symptoms from cases over controls *inflates the odds ratio*, because $p(S+|D-)$ will be underestimated to a greater extent than $p(S+|D+)$. Therefore, the results strongly suggest that, in CPRD studies restricting analysis to coded records, the association between recognised alarm symptoms and cancer will be inflated by biased estimates of the odds ratio. For example, in the bladder cancer study, the biased odds ratio for cancer in patients following a single episode of visible haematuria was 100.5 (95%CI:

78.1–129.3). After correcting for recording style bias the odds ratio was estimated at 75.9 (60.9–94.6) (Table 10.6).

In contrast, the association between ‘low-risk but not no-risk’ symptoms and cancer, as assessed by the odds ratio, is unlikely to be unaffected by recording style bias. Indeed, this was the case for abdominal pain in the bladder cancer study, where the odds ratio of 2.1 (95% CI: 1.9–2.4) was unaltered by the addition of text records (2.2, 95% CI: 2.0–2.4) (Table 10.7).

The effect of recording style bias in terms of modelling the symptomatic presentation of cancer in the original studies is discussed below for univariable and multivariable analysis separately.[4,5]

13.4.3.1 In univariable analysis

Discussion of the loss of symptom information in the hidden text generally and the resulting failure of a symptom to meet the pre-specified (though arguably arbitrary) 5% threshold for inclusion in univariable analysis was discussed separately above (see Section 13.3.7.1).

Symptoms meeting the 5% threshold for inclusion in univariable analysis level were assessed individually as to whether they had an independent association with cancer (z test of whether the odds ratio differs significantly from 1). My study shows that recording style bias either has no effect on (for ‘low-risk but not no-risk’ symptoms) or inflates (recognised alarm symptoms) the odds ratio. Furthermore, to ensure that important variables were not omitted, the p value for retention on the z test was set to 0.1. Therefore, researchers can be

reassured that recording style bias, even if present, has little effect at the univariable analysis stage.

13.4.3.2 In multivariable analysis

In contrast, researchers in the field of cancer diagnostics should be aware that recording style bias may have an impact on the final selection of which symptoms best characterise the presentation of the cancer being investigated. This is because the modelling method entails using sequential regressions at the multivariable analysis stage to select those features with the strongest association with cancer. Recording style bias favours the retention of recognised alarm symptoms over ‘low-risk but not no-risk’ symptoms, because it selectively inflates the odds ratios of the former but not the latter. A ‘low-risk but not no-risk’ symptom would be falsely rejected if its odds ratio (an accurate estimate) were caused to fall below the selected p-value for retention indirectly, through the inflated odds ratio of an alarm symptom. Strictly, this would only happen if loss of records of attendance for the alarm and ‘low-risk but not no-risk’ symptoms in the hidden text predominantly affected the same cases. If this requirement were satisfied, the association between the low-risk symptom and cancer could fall to a level low enough that the symptom were rejected from the model. It is unlikely this happened, given that the proportion of cases whose entire records for a high-risk symptom were lost in the hidden text was in the order of 20% (visible haematuria in bladder cancer) to 30% (jaundice in pancreatic cancer). Therefore, this gives some reassurance that recording style bias between cases and controls has rarely led to the inappropriate omission of

symptoms from a final model of the clinical features of cancer.[4,5] Indeed, the CAPER studies can, and do, check the omitted variables against the final model, to confirm that their exclusion is appropriate.

13.4.3.3 Summary

To summarise, recording style bias is likely to inflate outcome measures for recognised alarm symptoms in CPRD studies of diseases strongly associated with that alarm symptom. In contrast, outcome measures for 'low-risk but not no-risk' symptoms are not affected by recording style bias.

Perversely, this may not always matter in cancer diagnostics; for example, even after correction for recording style bias, the PPV for jaundice in pancreatic cancer still considerably exceeded the current 3% PPV threshold for referral for investigation (reducing from 12.77%, 95% CI: 7.25–21.58% to 6.25%, 4.46–8.68%).[5]

A more important concern is that recording style bias potentially introduces a positive feedback effect between research and recording practice. I propose that the selective inflation of risk estimates for alarm symptoms has two effects. First, it reinforces clinicians' views of the (already known) importance of these symptoms, increasing the chances that they will be recorded using a code. Secondly, it further marginalises clinicians' views of the importance of 'low-risk but not no-risk' symptoms, increasing the chances that their recording is relegated to the hidden text. The net effect is to increase recording style bias still further.

The discussion now moves on to exploring the potential impact of recording style bias beyond the CPRD.

13.4.4 Impact of recording style bias for NICE guidance for suspected cancer

In the recent update of NICE guidance,[9] one of the original studies extended by this PhD – Stapley *et al* (2012) – was the sole source used to assess evidence of pancreatic cancer in patients with jaundice.[5]

Regarding the evidence of bladder cancer in patients with visible haematuria, the NICE committee decided to omit Shephard *et al* from its meta-analysis because of its case–control design.[4]

The potential impact of loss of records in the hidden text and associated bias of the positive predictive value (PPV) of visible haematuria for bladder cancer in the five studies[6,66,67,68,69] in the NICE meta-analysis[65] is discussed below.

It should be noted that direct comparison of my PPV values with the individual values reported in the five studies in the NICE meta-analysis is limited by between-study differences in disease prevalence.[159] This is not just because PPV itself is dependent on disease prevalence, but also because some of the studies were conducted before and some after the change in definition of bladder cancer in 1998.[62] Therefore, I limit the discussion to the direction of change that would be observed if correction of recording style bias were made.

13.4.4.1 Jones *et al* – CPRD study

Jones *et al*[69] carried out a cohort study using coded CPRD data. Therefore, it will be vulnerable to biased recording of visible haematuria, favouring identification of the symptom in patients who receive a diagnosis of bladder cancer compared with healthy patients in the cohort.

As to the potential effect of recording style bias on the positive predictive value (PPV), it is first necessary to review how this was calculated. As it was a cohort design, the PPV was estimated arithmetically in a contingency table and quoted for a particular cancer incidence. As described in Section 6.12.3.5, the PPV is the number of patients with visible haematuria who also had cancer (i.e. the 'true-positives') divided by the total number of patients who had visible haematuria, whether they had cancer or not (i.e. the sum of true-positives and false-positives). Loss of records of attendance for visible haematuria in the hidden text will lead to underestimation of both the numbers of true-positives and false-positives. Recording style bias similar to that observed in my study will inflate the PPV, because the number of false-positives will be underestimated to a greater extent than the number of true-positives.

I suggest that the degree of recording style bias for visible haematuria in Jones' cohort study is greater than in my case-control study, despite the fact that both were conducted using CPRD data. This is related to differences in study design and consequently the characteristics of the patients recruited to each study.

My study had a case–control design, in which cases were matched to controls on age, sex and GP practice. The increased incidence of bladder cancer in men compared with women led to a gender imbalance of 19,017/26,633 (71%) men and 7,618/26,633 (29%) women in my study. In contrast, Jones followed a cohort design and selected all 923,605 patients who were registered with a CPRD practice between 1 January and 31 December 1994. While the gender balance is not reported, it is likely to be roughly 50% male and 50% female.

My results indicate that recording style bias for visible haematuria is more marked in women than in men. This resulted in a greatly overestimated PPV for bladder cancer (pre-1998 definition) in women with bladder cancer of 5.10% (95% CI: 3.30–7.80%), which was corrected to 2.17% (95% CI: 1.68–2.81%) after hidden records for the female controls were unmasked by examination of the text records (see Table 10.3). In contrast, the overestimation of the PPV for bladder cancer in men was less marked. The coded records suggested a value of 5.14% (95% CI: 4.46–5.93%), which was 'corrected' down to 3.97% (95% CI: 3.54–4.46%) by addition of text-only records.

This was further supported by my *post-hoc* analysis, which suggested that when the GP suspected that the visible haematuria had a benign cause, they were more likely to use the free text than a code to record the symptom. This effect was much greater for female than for male patients (odds ratio of text-only recording was 9.3, 95%CI: 5.0–17.4, $p<0.0001$ for women, and 3.0, 95% CI: 2.2–4.0, $p<0.0001$ for men).

My study yielded a biased PPV of visible haematuria for bladder cancer in all patients over 40 years old of 3.98% (95% CI: 3.47–4.57), which was 'corrected' to 2.85% (95% CI: 2.57–3.17%) after addition of text-only records. This suggests a correction factor for recording style bias of 1.4 (Table 10.2) in my case–control study. The comparatively large number of healthy female control patients in the Jones *et al* cohort study means that this correction factor will be unable to correct the attendant recording style bias fully. Therefore, his reported PPV in all patients of 4.2% (95% CI: 3.8–4.6%) is likely to be considerably inflated.

13.4.4.2 In the QResearch® study

Hippisley-Cox *et al* carried out a cohort study using the QResearch® database of electronic medical records. The aim was to derive an algorithm for assessing the risk of urinary tract cancer in patients presenting to primary care, named QCancer® (Renal).[6]

13.4.4.2.1 Potential for loss of records in hidden text in QResearch®

QResearch® is based on medical records obtained in every-day general practice using what was then known as Egton Medical Information Systems (EMIS) software.[a] In terms of recording clinical information, EMIS and ViSion (the software used by the majority of practices that contribute to the CRPD) differ fundamentally. ViSion requires GPs to select an appropriate Read code

[a] Egton Medical Information Systems is now known as EMIS Health.

before they can access the free text comments box. In contrast, in EMIS GPs start recording the consultation by typing in the text box, and then have the option to select Read codes that are suggested based on character matches with the free text the GP is typing. So in ViSion every consultation must be accompanied by at least one Read code, while in EMIS the entire clinical encounter can be recorded in the hidden text.

Clinical events included in the Quality and Outcomes Framework (QoF) (see Section 13.3.6.5) are likely to be recorded using a code in EMIS. However, for anything outside QoF, there is little incentive for GPs to break their train of thought mid-typing in order to select a suggested Read code. Indeed, one could argue that GPs who are unable to touch-type may even be unaware of the range of Read code choices on offer, because their attention will be focused on the keyboard rather than the screen.

For these reasons, the Hippisley-Cox study is likely to have underestimated the prevalence of visible haematuria (which is not covered by QoF) reported by her cohort of patients, regardless of whether they receive a cancer diagnosis. Evidence of this would suggest that studies based in EMIS are, at the very least, more vulnerable than CPRD studies to *loss of data to research* through use of the hidden text to record clinical events. Greater underestimation of the frequency of symptoms will not affect recording style bias if the excess loss of symptom information due to EMIS' ready access to text fields is the same in patients with cancer as in the healthy cohort, but it was not possible to test this in my study.

13.4.4.2.2 Frequency of visible haematuria

Testing for excess loss of data in QResearch® compared with CPRD studies would require a direct comparison of Hippisley-Cox's estimate of the frequency of visible haematuria with that obtained in my study. This is complicated by differences in our study design. On the one hand, mine is a case–control study of 26,633 patients over 40 years of age, resulting in a population that is artificially enriched with bladder cancer patients.[61] On the other hand, the Hippisley-Cox study studied over 3.5 million patients, who, at the time of the study, were between 30 and 84 years old and did not have a renal tract cancer diagnosis.

Hippisley-Cox *et al* reported a crude incidence rate of 298 per 100,000 person-years. These figures suggest that a maximum of 0.3% of patients in the whole cohort (including healthy patients and those who received a renal tract cancer diagnosis during the study) experienced an episode of haematuria. This ignores the possibility of repeat attendance by individual patients, which would lower the figures still further.

The symptom thesaurus for visible haematuria used in my study and by Shephard *et al*[4] identified that, in our healthy controls alone, 196/21,718 (0.9%) had visible haematuria.

These findings support the hypothesis that studies based on coded records from EMIS underestimate the frequency of symptoms through increased use of the hidden text to records symptoms. At the very least, the increased number of

'false-negatives' (i.e. cancer patients whose symptoms were recorded solely in the text) will lead to the misleading conclusion that too high a percentage of cases are asymptomatic.

13.4.4.2.3 Study population factors

As discussed above in relation to Jones *et al*[69] (see Section 13.4.4.1), my analysis shows that women are three times more likely than men to have their visible haematuria recorded solely in the text when the GP suspects that there is a benign cause. In contrast to my study, which had a gender imbalance of approximately 2.5 men to every woman, Hippisley-Cox *et al* used a cohort design and recruited more than 3 million patients with *similar* numbers of men and women. Therefore, compared with my case–control study, the Hippisley-Cox *et al* study had a far greater proportion of control healthy patients who were women.

This suggests that the Hippisley-Cox study was highly vulnerable to recording style bias that favoured identification of visible haematuria in the patients who were diagnosed with bladder cancer. This would lead to biased estimates of the PPV, as is discussed below.

13.4.4.2.4 Positive predictive value

The Hippisley-Cox cohort study reported a PPV for bladder cancer of 6.48% (95% CI: 6.1–6.8%) at an incidence rate of 70 per 100,000 person-years for all urological cancers (of which 79% were bladder cancer). This value was at the

upper end of the PPV estimated in the meta-analysis of five studies considered by NICE (5.1%, 95% CI: 3.2–8.0%).

I suggest that two factors are acting in the Hippisley-Cox *et al* study to inflate this estimate considerably above the true value. First and foremost, the cohort design results in the inclusion of a high proportion of healthy female controls, whose visible haematuria is three times more likely to be recorded solely in the text compared with healthy male controls. Secondly, it is feasible that the recording style bias in QResearch® studies is even greater than that in CPRD studies, owing to the GPs' easy access to text fields.

13.4.4.3 In The Health Improvement Network (THIN) study

The Collins and Altman[67] study was cohort in design, carried out using data from the THIN database of electronic medical records to validate Hippisley-Cox's QCancer® (Renal) study discussed above. Practices that contribute data to THIN share the same IT system as the CPRD, namely ViSion. Therefore, the Collins and Altman study is likely to be affected by recording style bias to the same extent as a CPRD study. Collins and Altman reported a PPV for bladder in patients with visible haematuria of 4.35% (95% CI: 4.1–4.6%), which again is likely to be an overestimate given the data's vulnerability to recording style bias.

Like Hippisley-Cox *et al*,[6] Collin and Altman[67] recruited similar numbers of men and women; therefore, the degree of recording style bias is likely to be even greater than that observed in my study, owing to their greater prevalence of healthy female controls.

13.4.4.4 In the Network of Sentinel General Practitioners in Belgium

The Bruyninckx *et al* study was carried out using primary care data from the Network of Sentinel General Practitioners in the Belgian Healthcare System.[66] Participants in the Bruyninckx *et al* study were those patients who had received a diagnosis of urological cancer in the period 1993–1994. Patients were asked specifically if they had complained to their GP about visible haematuria prior to their diagnosis. Therefore, while the study is vulnerable to recall bias, the estimates of PPV are not likely to be inflated by the recording style bias identified in my thesis.

13.4.4.5 In the Vanderbilt University Medical Center Research Derivative

Friedlander *et al* was a billing study was conducted using data from the Vanderbilt University Medical Center Research Derivative.[68] This is a database containing complete administrative and clinical information about every patient treated in the Vanderbilt Health System (a private healthcare facility in Tennessee, USA), in which coding is strongly driven by billing.

Patient data were managed using a platform called REDCap (Research Electronic Data Capture), which strongly discourages the use of free text fields (confirming that they do exist) as their content is difficult to analyse.[160] The presence of haematuria was confirmed either by urinalysis (for which coding is obligatory), or by the presence of a diagnostic code. The classification of haematuria as a diagnosis means that it would almost certainly always be

recorded as a code, for billing purposes. Therefore, it is highly unlikely that recording style bias affected the outcomes measures.

13.4.4.6 Summary

The NICE meta-analysis reported a summary PPV of visible haematuria for bladder or renal cancer in patients aged 15–100 years old of 5.1% (95% CI: 3.2–8.0%).[65] Three[6,67,69] of the five studies used in this meta-analysis[65] are vulnerable to recording style bias that inflates estimates of PPV. The recording style bias in these three studies is likely to be at least at the level observed in my study, and probably greater owing to their cohort design. This relates to the three-fold propensity of GPs to record visible haematuria in the hidden text when female (rather than male) patients present and have what the GP suspects is a benign cause of their symptom. The cohort studies, which recruited equal numbers of men and women, were even more vulnerable to recording style bias than my case–control design, which had a male : female ratio of 2.5 : 1.

13.4.5 Comparison with existing literature

Just one study, to the best of my knowledge, has analysed the content of text fields in the CPRD to identify how much symptom information is stored there and not as codes.[33] This work was reported in conference proceedings and has not been published in a peer-reviewed journal. I have identified no CPRD studies that report on differential recording styles in comparison groups or on

the association between use of codes vs text and perceived symptom significance.

13.5 Discussion of clinical findings: the risk of bladder cancer in patients with non-visible haematuria

My study is the first report of the risk of bladder cancer in primary care patients with non-visible haematuria. For the pre-1998 definition of bladder cancer,[a] the risk was 1.60% (95% CI: 1.22–2.10%) for all patients of 40 years or older (see Table 10.2). This is just over one-half the risk estimated in patients with visible haematuria, after correction for recording style bias (2.85%, 95% CI: 2.57–3.17%), and more than twice that of the next highest risk symptom – dysuria (0.66%, 95% CI: 0.56–0.78%).

My estimation of the risk of bladder cancer in patients with non-visible haematuria was most timely, as it coincided with revision of the guidelines on referral for suspected cancer by The National Institute for Health and Care Excellence (NICE). As a consequence, my risk estimates for the post-1998 definition of bladder cancer[b] were used by NICE. The data in the over-60s were particularly pertinent, putting the risk of bladder cancer at 1.66% (95% CI: 1.22–

[a] Including carcinoma *in situ* of bladder (ICD10 code D09.0) and neoplasm of uncertain behaviour of bladder (D41.4).

[b] ICD10 code C67, excluding carcinoma *in situ* of bladder (ICD10 code D09.0) and neoplasm of uncertain behaviour of bladder (D41.4).

2.26%) (see Table 16.2). This is approximately half the risk estimated in patients with visible haematuria (3.03%, 95% CI: 2.67–3.44%, see Table 16.2), but much greater than that in patients presenting with any of the single features of bladder cancer identified. Dysuria was the next highest risk symptom in the over-60s, with a positive predictive value of 0.61% (95% CI: 0.50–0.74%).

In light of these findings, NICE included a new recommendation to refer patients aged over 60 who have unexplained non-visible haematuria and either dysuria or a raised white cell count on a blood test.[9]

There are no comparison studies estimating the risk of bladder cancer in patients presenting to primary care with non-visible haematuria. Some studies have estimated the positive predictive value of non-visible haematuria for urological malignancy in a secondary care setting. For example, Edwards *et al* conducted a prospective analysis of 4,020 patients attending a haematuria clinic in the UK between October 1998 and August 2003. They identified 94 malignancies (renal cell carcinoma and transitional cell carcinoma) in 1,949 patients with non-visible haematuria, yielding a positive predictive value of 4.8% for all age groups combined.[161] In the UK, however, not all patients with non-visible haematuria are referred for investigation; therefore, studies conducted in secondary care are not generalisable to the primary care setting, owing to selection bias. Indeed, Edwards *et al* recruited patients from a haematuria clinic, to which referral is only made after urinary tract infection has been excluded.

A limitation of the study is the detection bias discussed in Section 13.2.5.3. This is likely to mean that I have underestimated the frequency of non-visible

haematuria in cases and controls. It is possible that this has inflated the positive predictive value, because the levels of underestimation are likely to be greater in controls than in cases.

14 Conclusion

This thesis has studied a neglected area of research; namely, the potential for systematic bias in studies of electronic medical records that restrict their analysis to codes, omitting anything recorded in the text.

To conclude, I address my original research questions.

14.1 How much symptom information is documented in electronic medical records using text rather than a code?

My study shows that a considerable amount of symptom information is documented in electronic medical records using text rather than a code. At the event level, the amount of text-only recording varied, from as much as 60% (298/494) for non-visible haematuria in bladder cancer to 31% (2,215/7,191) for abdominal pain in pancreatic cancer.

Use of text recording persists with repeat visits to the GP such that, for some patients, their entire record of attendance for the symptom is lost in the hidden text. In numerical terms, non-visible haematuria was shown to be affected the most (219/372=59%); visible haematuria, the least (696/3,483=19%).

Therefore, studies restricting their analysis to coded records underestimate both the frequency of attendance for symptoms, and the number of affected patients.

14.2 Are studies of coded data vulnerable to bias arising from the differential use of text and codes between comparison groups?

The answer to this is a qualified yes. GPs have an increased tendency to code attendance for alarm symptoms in those patients later diagnosed with a cancer that is strongly associated with the symptom. This leads to recording style bias that affects studies whose analysis is restricted to codes. This is because codes detect more complete information about symptoms from the cases than from the controls. This inflates not only the measures of association between *recognised alarm symptoms* and cancer, but also the risk estimates for cancer.

14.3 Does recording style vary with type of symptom?

The answer to this depends on whether you are looking at the event or the patient level.

At the event level, there is no evidence of an association between recording style and the type of symptom.

At the patient level, however, an association becomes apparent. As just discussed, GPs are more likely to code alarm symptoms than use the text when the alarm symptoms are presented by patients later diagnosed with a strongly associated cancer.

In contrast, for features that do not have a strong association with malignant pathology, GPs' recording style is similar whether the patient is later diagnosed

with cancer or remains free of the disease. Therefore, studies restricting their analysis to codes will produce unbiased measures of association between these features and cancer, and unbiased risk estimates.

14.4 Does the recording style vary with the clinical context of a symptom's presentation?

The answer to this is yes. As described above, GPs have an increased tendency to code attendance for alarm symptoms in patients later diagnosed with a cancer that is strongly associated with the symptom. However, this strong preference for coding is not retained for alarm symptoms presented by patients later diagnosed with an unconnected cancer. This suggests that GPs tend to code attendance for alarm symptoms *when they suspect* that the cause is an underlying malignancy.

The patient's gender affects the recording style of visible haematuria in particular. GPs are more likely to use the text to record attendance for visible haematuria for female than for male patients where they suspect that the underlying cause is benign.

14.5 Do text data provide additional value to coded data?

The answer to this is yes, on a number of grounds. First, including text records permits a more accurate estimation of both the frequency of attendance for a symptom, and the number of patients affected. Including text records enabled

the first estimate of the risk of bladder cancer in patients presenting with non-visible haematuria[8] – a direct illustration of their added value.

Secondly, studies that include text data produce: unbiased measures of association between alarm symptoms and cancer; unbiased risk estimates of cancer in symptomatic patients; and unbiased estimates of the timing of first symptom presentation.

14.6 Summary

Inclusion of text records increases the accuracy of outcome measures in studies of observational data from electronic medical records. This is important because electronic medical records are increasingly used in epidemiological research, and provide much of the evidence on which national guidelines are based.

Regrettably, it is no longer permissible for the CPRD to collect text records made by GPs. Therefore, future studies of CPRD records will be limited to anticipating the size and direction of the recording style in their data. My study suggests that researchers should consider the following factors when assessing their study's vulnerability to recording style bias.

First, researchers should identify whether the symptom is included as part of the Quality and Outcomes Framework. If it is, the researcher can be confident that symptom occurrence will be coded consistently for all patients, with minimal recording style bias.

Second, for symptoms that are not included in the Quality and Outcomes Framework, researchers should consider the symptom's clinical significance. Our findings suggest that GPs preferentially code clinical features they consider significant to a diagnosis, while tending to use hidden text to record those that they think are not.

For alarm symptoms, researchers should anticipate recording style bias that favours detection of a history of the symptom in patients later diagnosed with a disease with which the symptom is strongly associated. In contrast, for 'low-risk but not no-risk' symptoms, recording style bias is likely to be minimal or absent. While the risk estimates for the latter symptoms are unaffected, compared with the inflated estimates for alarm symptoms, they appear to be relatively low and unimportant, which inappropriately marginalises them in the clinicians' eyes.

Therefore, recording style bias introduces a positive feedback loop between research and clinical practice. The inflated risk estimates for symptoms reinforces clinicians' views of the symptom's importance, further increasing the amount of recording style bias introduced by GPs when they make the medical record.

I understand that the CPRD is lobbying hard to overturn the current moratorium on the collection and availability of text data for research. I support them in their endeavours because, as this thesis shows, text records are a rich and valuable resource that improves the accuracy of epidemiological research.

15 Appendices

15.1 Appendix 1: Literature search tables

Table 15.1 The search terms and number and details of references found. The Boolean/Phrase search mode was chosen and the study period was 1946 to May 2014

Search terms	Number and details of references found
'General Practice Research Database' AND 'unstructured text'	3[28,118,162]
'General Practice Research Database' AND 'free text'	11[1,22,23,24,28,30,31,118,162,163,164]
'GPRD' AND 'free text'	7[1,22,23,24,28,162,163]
'GPRD' AND 'unstructured text'	2[28,162]
'free text' AND 'electronic health records' AND 'UK'	9[13,15,28,32,165,166,167,168,169]
'free text' AND 'electronic medical record' AND 'UK'	0
'unstructured text' AND 'electronic health records' AND 'UK'	1[28]
'unstructured text' AND 'electronic medical record' AND 'UK'	0
'uncoded text' AND 'electronic health records' AND 'UK'	0
'uncoded text' AND 'electronic medical record' AND 'UK'	0
'Clinical Practice Research Datalink' AND 'validation'	7[25,27,170,171,172,173,174]
'General Practice Research Database' AND 'validation'	28[10,17,20,21,22,23,25,26,175,176,177,178,179,180,181,182,183,184,185,186,187,188,189,190,191,192,193,194]

Table 15.2 Reasons for study exclusion

Reason	Study
Uncoded data not examined	10,164,165,170,171,172,173,174,175,176,177,178,179,180,181,182,183,184,185,186,188,189,190,191,192,193,194
Not GPRD/CPRD	17,167,169,187
Used to develop natural language processing tool	1,118,162,168
Publication unavailable	163
Protocol for systematic review	166

15.2 Appendix 2: Disease thesauri

Table 15.3 Bladder cancer (cancer site 1) disease thesaurus codes

Medcode	Description	Read code
779	Malignant neoplasm of urinary bladder	B49..00
6436	[M]Transitional cell carcinoma NOS	BB43.00
7187	Carcinoma *in situ* of bladder	B837.00
9712	[M]Papillary transitional cell carcinoma	BB4A.00
16926	Neoplasm of unspecified nature of bladder	BA04.00
19162	Malignant neoplasm of anterior wall of urinary bladder	B493.00
21652	[M]Transitional cell carcinoma *in situ*	BB42.00
22146	Secondary malignant neoplasm of bladder	B581100
28241	Malignant neoplasm of ureteric orifice	B496.00
31102	Malignant neoplasm of urinary bladder NOS	B49z.00
35963	Malignant neoplasm of lateral wall of urinary bladder	B492.00
36949	Malignant neoplasm of other site of urinary bladder	B49y.00
38862	Malignant neoplasm of trigone of urinary bladder	B490.00
41571	Malignant neoplasm of bladder neck	B495.00
42012	Malignant neoplasm of posterior wall of urinary bladder	B494.00
42023	Malignant neoplasm of urachus	B497.00
44996	Malignant neoplasm of dome of urinary bladder	B491.00
47801	Malignant neoplasm, overlapping lesion of bladder	B49y000
58798	[M]Transitional cell carcinoma, spindle cell type	BB47.00
97091	[X]2ndry malignant neoplasm/bladder+oth+unsp urinary organs	ByuC500

Table 15.4 Pancreatic cancer (cancer site 12) disease thesaurus codes

Medcode	Description	Read code
8166	Malignant neoplasm of pancreas	B17..00
8771	Malignant neoplasm of head of pancreas	B170.00
9224	[M]Insulinoma NOS	BB5B200
10949	Malignant neoplasm of ampulla of Vater	B162.00
16931	Carcinoma *in situ* of pancreas	B80z000
21792	Carcinoma *in situ* of ampulla of Vater	B808600
26858	[M]Gastrinoma and carcinomas	BB5C.00
32294	[M]Glucagonoma, malignant	BB5B500
34388	Malignant neoplasm of pancreas NOS	B17z.00
35535	Malignant neoplasm of pancreatic duct	B173.00
35718	[M]Gastrinoma NOS	BB5C000
35795	Malignant neoplasm of Islets of Langerhans	B174.00
39870	Malignant neoplasm of tail of pancreas	B172.00
40810	Malignant neoplasm of body of pancreas	B171.00
43594	[M]Gastrinoma or carcinoma NOS	BB5Cz00
48537	Malignant neoplasm of other specified sites of pancreas	B17y.00
49629	[M]Gastrinoma, malignant	BB5C100
55675	Endocrine tumour of pancreas	B717011
58022	[M]Glucagonoma NOS	BB5B400
63102	[M]Islet cell carcinoma	BB5B100
95609	[M]Insulinoma, malignant	BB5B300
95783	Malignant neoplasm of specified site of pancreas NOS	B17yz00
96635	Malignant neoplasm of ectopic pancreatic tissue	B17y000

Medcode	Description	Read code
97875	Malignant neoplasm, overlapping lesion of pancreas	B175.00
98825	[M]Mixed islet cell and exocrine adenocarcinoma	BB5B600

15.3 Appendix 3: Algorithm construction

15.3.1 Grammar – a quick tour

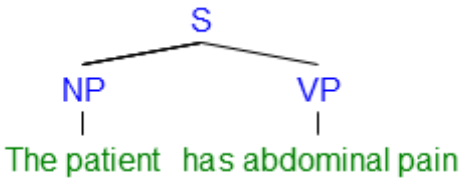
15.3.1.1 Introduction

A sentence is defined as consisting of a subject and predicate, where the predicate contains the verb and gives information about the subject. The number of grammatically correct sentences that can be formed from a group of words is far less than the number of possible combinations. For example, there are 5! (= 120) ways of arranging the five words ‘man’, ‘ball’, ‘a’, ‘the’ and ‘kicked’; however, only six of these form grammatically correct sentences and not all of these are meaningful.[195] Plainly then there must be rules governing how words can be arranged to form phrases, clauses and sentences.

15.3.1.2 Constituent phrases of a sentence

In linguistics, the set of rules governing how words can be arranged grammatically in a sentence is called **syntax**. Syntactically, sentences are thought to consist of a **noun phrase** (defined as a group of words that behave as a noun) and a **verb phrase** (defined as the main verb and its auxiliaries) (see Figure 15.1).

Figure 15.1 The basic structure of a sentence (S) can be shown using a syntax tree diagram. Abbreviations: NP, noun phrase; VP, verb phrase. (Image generated using Syntax Tree Generator Copyright © 2011 by Miles Shang **mail@mshang.ca**)

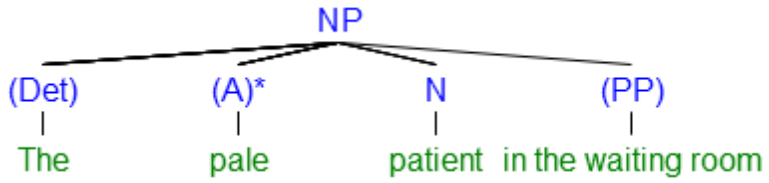


Delving deeper into syntax gives a framework on which to build an algorithm that can interpret and assign a particular meaning to phrases, clauses and sentences.

15.3.1.2.1 Noun phrase

As Figure 15.2 shows, the constituent parts of a noun phrase are the determiner, adjective, noun and prepositional phrase, of which only the noun is obligatory. A determiner is a word used before a noun to show which particular example of the noun is meant; for example, articles (e.g. the), possessives (e.g. my), quantifiers (e.g. no) and demonstratives (e.g. these). A prepositional phrase is one that both follows and modifies the noun.

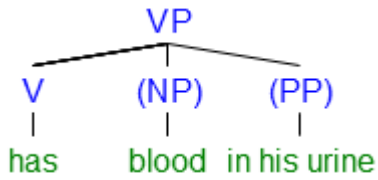
Figure 15.2 A syntax tree can be used to show the constituent parts of a noun phrase. Abbreviations: NP, noun phrase; Det, determiner; A, adjective; N, noun; PP, prepositional phrase. Parentheses enclose those parts of speech that are optional and an asterisk indicates where there is no upper limit for the number of words that can be included in the sentence. (Image generated using Syntax Tree Generator Copyright © 2011 by Miles Shang mail@mshang.ca)



15.3.1.2.2 Verb phrase

The constituent parts of the verb phrase, shown in Figure 15.3, are the verb (obligatory), noun phrase and a modifying prepositional phrase.

Figure 15.3 A syntax tree can be used to show the constituent parts of a verb phrase. Abbreviations: VP, verb phrase; V, verb; NP, noun phrase; PP, prepositional phrase. Parentheses enclose those parts of speech that are optional. (Image generated using Syntax Tree Generator Copyright © 2011 by Miles Shang mail@mshang.ca)



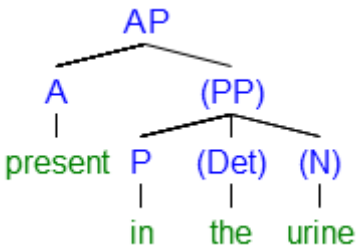
Verbs are classed as either auxiliary or main, where the auxiliary form determines the mood, tense or aspect the main verb that follows. For example, in the sentence ‘The patient bleeds’, ‘bleeds’ is a main verb; whereas, in the

sentence ‘The patient does bleed’, ‘does’ is an example of an auxiliary verb determining the mood of main verb ‘bleed’. Auxiliary verbs are particularly important when it comes to reversing the meaning of a phrase, as explained below.

15.3.1.2.3 Adjective and adverb phrases

Sentences can also include optional adjectives (or adjective phrases). As shown in Figure 15.4, the adjective phrase consists of an obligatory adjective, which can be modified by a prepositional phrase.

Figure 15.4 A syntax tree can be used to show the constituent parts of an adjective phrase. Abbreviations: AP, adjective phrase; A, adjective; PP, prepositional phrase; P preposition; Det, determiner; N, noun. Only the adjective is obligatory. (Image generated using Syntax Tree Generator Copyright © 2011 by Miles Shang **mail@mshang.ca**)



Finally, adverbs impart information about their associated verbs, adjectives or other adverbs.

Negation is the denial of the truth of a clause or statement. From the figures above, I worked out where negation can fall within the sentence, and used these rules to generate the search terms of my algorithm. As shown in Figure

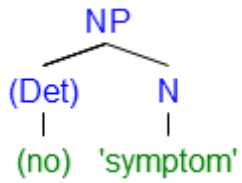
15.1, basic sentence construction involves a noun phrase and a verb phrase. Bearing in mind that one negative will negate another, the negation can be placed in one or other of these phrases and not both at the same time.

15.3.1.3 Negation of the noun phrase

15.3.1.3.1 Using the determiner

A noun phrase consists of an obligatory noun plus the optional determiner, adjective(s) and prepositional phrase (Figure 15.2). In my algorithm, I was looking to identify negative reports of symptoms in patients; therefore, the obligatory noun in the noun phrase will be either the **symptoms** themselves or the **patient**. The determiner also allows the introduction of negation, through the use of words such as ‘no’ and ‘none’ (Figure 15.5):

Figure 15.5 Negation of a noun phrase is introduced by the determiner. (Image generated using Syntax Tree Generator Copyright © 2011 by Miles Shang **mail@mshang.ca**)



The text strings provided by the CPRD are of one-to-one consultations between a patient and their GP, during which it is unlikely that the GP will negate the patient (i.e. no patient…); therefore, I thought it reasonable to ignore this option.

This gives me my first rule for the algorithm:

Algorithm rule 1

negative determiners placed before the symptom, for example

- ‘no’ + ‘symptom’
- ‘nil’ + ‘symptom’
- ‘no more’ + ‘symptom’[a]
- ‘no further’ + ‘symptom’[b]

15.3.1.3.2 Using affixes

Combining nouns with the affixes ‘un-‘, ‘a-‘, ‘de-‘, ‘dis-‘, ‘in-‘, ‘-less’ and ‘mis-’ will reverse their sense. In this study I was searching for the following symptoms: haematuria, jaundice, icterus and abdominal pain, none of which have negated forms created using the above affixes. Therefore, this rule has no application here, but it may be useful in others so I’ll include it:

[a] This form indicates that the patient has experienced the symptom in the past, but is free of it at the time of consultation.

[b] This form indicates that the patient has experienced the symptom in the past, but is free of it at the time of consultation.

Algorithm rule 2

affixes negating symptoms, for example**:**

- amenorrhoea

15.3.1.3.3 Using the preposition

Words such as ‘without’ can be used before the symptom to indicate their absence, giving me rule 3:

Algorithm rule 3

negative prepositions negating the symptom, for example:

- without + (any) + ‘symptom’

15.3.1.4 Negation of the verb phrase

A verb phrase consists of an obligatory verb, and optional noun and prepositional phrases (Figure 15.3). Negation of a verb can take several forms, which are described below.

15.3.1.4.1 Using 'not'

A commonly used construction is the negative adverb '**not**', which reverses the sense of a positive verb.[196] It is important to remember that verbs are classified as auxiliary and/or main (see Section 15.3.1.2.2).

1. **Auxiliary verbs**: 'Not' sandwiched between auxiliary and main verbs reverses the meaning of the latter[112] (p 15) [195] (p 152); for example, 'The patient does not present with blood in his urine'. 'Not' must be used with auxiliary verbs, so the opposite of this is 'The patient presents with blood in his urine'. In other words, negation cannot take the form 'The patient presents not with blood in his urine'. This gives me my fourth rule:

Algorithm rule 4

'auxiliary verb' + 'not' + 'main verb' + (any) + 'symptom', for example:

- 'not' + 'have' + (any) + 'symptom'
- 'not' + 'got' + (any) + 'symptom'
- 'not' + 'complain(ing)/(ed)' + 'of' + (any) + 'symptom'
- 'not' + 'present(ing)/(ed)' + 'with' + (any) + 'symptom'
- 'not' + 'encounter(ed)' + (any) + 'symptom'
- 'not' + 'exhibit(ed)' + (any) + 'symptom'
- 'not' + 'notice(d)' + (any) + 'symptom'
- 'not' + 'reveal' + (any) + 'symptom'
- 'not' + 'see(n)' + 'symptom'
- 'not' + 'suggest' + (any) + 'symptom'

---

2. **Infinitive verbs** [195] (p 163): '**not**' can be placed before an infinitive to negate its meaning, giving me rule 5:

Algorithm rule 5

'not' + 'infinitive verb' + 'symptom', for example:

- 'not' + 'to be' + 'symptom'

15.3.1.4.2 Using 'no longer'[a]

Algorithm rule 6

'noun phrase' + 'negative adjective' + 'verb phrase', for example:

- patient + no longer + has + (any)+ 'symptom'
- patient + no longer + complains of + (any) + 'symptom'

15.3.1.4.3 Using negative verbs

Where a verb has a negative meaning, no other form of negation is required. This is most likely to occur when the patient is the noun (rule 7a), although it

[a] This form indicates that the patient has experienced the symptom in the past, but is free of it at the time of consultation.

can be used in the present perfect tense when the symptom is the noun (rule 7b):

Algorithm rule 7a

‘patient’ + ‘negative verb’ + ‘symptom’, for example:

- patient denies + (any)
- patient denied + (any)
- patient refutes + (any)
- patient refuted + (any)

Algorithm rule 7b

‘symptom’ + ‘negative verb’, for example:

- haematuria has ceased/stopped/cleared/resolved/gone/settled[a]

15.3.1.5 Negation of the adjective phrase

15.3.1.5.1 Using ‘not’, ‘never’, ‘no longer’

When ‘not’, ‘never’ or ‘no longer’ is placed before a positive adjective it negates it, giving me my eighth rule:

[a] This form indicates that the patient has experienced the symptom in the past, but is free of it at the time of consultation.

Algorithm rule 8

‘noun phrase’ + ‘is’ + ‘not’ + ‘adjective, for example:

- ‘symptom’ + ‘(is)’ + ‘not’ + present
- ‘symptom’ + ‘(is) + ‘no longer’ + present[a]
- ‘symptom’ + ‘(is)’ + ‘not’ + applicable
- ‘symptom’ + ‘(is)’ + ‘not’ + indicated

Note: in notes, doctors may omit the verb here and just write ‘haematuria not present’, for example.

15.3.1.5.2 Using negative adjectives

Searching for negative adjectives used to describe a symptom gives me my ninth rule:

Algorithm rule 9

‘noun phrase’ + ‘is’ + ‘negative adjective’, for example:

[a] This form indicates that the patient has experienced the symptom in the past, but is free of it at the time of consultation.

- ‘symptom’ + ‘(is)’+ absent/ruled out/negative (-ve)/ excluded/ rejected/ denied
- ‘patient’ + ‘(is)’ + asymptomatic/free of

15.3.1.6 Word patterns that introduce ambiguity

1. **Any**: a determiner used before a noun that is defined ‘as some, or even the smallest amount or number of’:

    a. ‘if any’ indicates an ambiguous observation and requires further investigation to determine the context.

Algorithm rule 11

‘if any’ + ‘symptom’

Classify these occurrences as ‘ambiguous and needing manual classification’

---

    b. ‘any XXX or symptom.’ (where XXX is another symptom) coming at the end of a string of text is ambiguous and requires further investigation to determine the context (caught by rule 12 below).

2. **Or**: a conjunction placed before a noun to connect different possibilities. It is also often used after a negative verb to mean neither one thing nor another,

which is the most likely use in this context. An alternative to 'or' is the solidus with variable spacing.

Algorithm rule 12

‘or’ + ‘symptom’

‘/’ + ‘symptom

Classify these occurrences as ‘ambiguous and needing manual classification’

---

3. **If**: conjunction to mean ‘that a particular thing can or will happen only after something else happens or becomes true’. When used before a symptom, 'if' indicates the possibility of that symptom and not a definite occurrence.

Algorithm rule 13

‘if’ + ‘symptom’

Classify these occurrences as ‘ambiguous and needing manual classification’

---

4. **Ago**: adverb meaning ‘back in time from the present’. This may mean that the symptom occurred in the past but is not a current concern. I deliberately sought the presence of this word in the text snippets and marked observations containing it as requiring manual assessment.

Algorithm rule 14

‘ago’

Classify these occurrences as ‘ambiguous and needing manual classification’

15.4 Appendix 4: Symptom thesauri

Table 15.5 Abdominal pain symptom thesauri

Medcode	Description	Thesaurus type:		
		Bladder	Pancreas	Generic
2056	[D] Abdominal colic	X	X	X
716	[D] Abdominal cramps	X		X
2383	Abdominal discomfort	X		
17762	[D] Abdominal migraine			X
2861	Abdominal migraine			X
6433	Abdominal migraine - symptom			X
1763	[D] Abdominal pain	X	X	X
177	Abdominal pain	X	X	X
3338	[D] Abdominal pain NOS	X	X	X
1976	Abdominal pain type	X	X	X
16402	[D] Abdominal tenderness	X		
29352	Abdominal wall pain	X	X	X
54385	[D] Acute abdomen			X
17586	Angina - abdominal			X
15908	Appendicular colic	X		X
930	Biliary colic			X
6285	Biliary colic			X
7306	Biliary colic symptom			X
4617	Central abdominal pain	X	X	X
1239	[D] Colic NOS			X
7812	Colicky abdominal pain	X		X
257	Dyspepsia		X	
542	[D] Epigastric pain		X	X
290	Epigastric pain		X	X
23756	[D] Evening colic	X		
8697	Flatulent dyspepsia		X	
28285	[D] Gas pain (abdominal)	X	X	X
11070	General abdominal pain-symptom	X	X	X
24661	Generalised abdominal pain	X	X	X
17324	Griping pain			X
1336	[D] Groin pain			X

Medcode	Description	Thesaurus type:		
		Bladder	Pancreas	Generic
628	[D] Hypochondrial pain			X
421	Iliac fossa pain	X		X
134	Indigestion		X	
7623	Indigestion NOS		X	
37118	Indigestion symptom NOS		X	
5862	Indigestion symptoms		X	
2982	Left iliac fossa pain			X
9061	[D] Left lower quadrant pain	X	X	X
8362	[D] Left upper quadrant pain	X	X	X
22608	Lower abdominal pain	X		X
5691	Non-colicky abdominal pain	X	X	X
19283	[D] Nonspecific abdominal pain	X	X	X
19223	O/E - abd. pain - epigastrium		X	X
42211	O/E - abd. pain - hypogastrium		X	X
21583	O/E - abd. pain - L.iliac			X
25630	O/E - abd. pain - L.lumbar			X
11647	O/E - abd. pain - R. iliac		X	X
36558	O/E - abd. pain - R.lumbar			X
24627	O/E - abd. pain - umbilical	X		X
37101	O/E - abd.pain-L.hypochondrium		X	X
12639	O/E - abd.pain-R.hypochondrium			X
15180	O/E - abdo. pain on palpation		X	X
5782	O/E - abdomen tender	X		
50590	O/E - abdominal rigidity	X		
73235	O/E - abdominal rigidity NOS	X		
20640	O/E - epigastric pain on palp.		X	X
42235	O/E - guarding - epigastrium			X
56084	O/E - guarding - hypogastrium			X
62933	O/E - guarding - umbilical	X		X
24584	O/E - guarding of abdomen			X
56094	O/E - guarding-L.hypochondrium			X
42195	O/E - guarding-R.hypochondrium			X
62934	O/E - rebound - epigastrium			X
62927	O/E - rebound - umbilical	X		X
17451	O/E - rebound tenderness			X
56091	O/E - rebound tenderness NOS			X
62965	O/E - rebound-R.hypochondrium			X
17636	O/E - umbilical pain on palp.	X		X
14916	O/E -abd.pain on palpation NOS		X	X

Medcode	Description	Thesaurus type:		
		Bladder	Pancreas	Generic
52402	[X] Other and unspecified abdominal pain		X	X
31062	[D] Other specified abdominal pain	X	X	X
16868	[D] Pain in left iliac fossa			X
16806	[D] Pain in right iliac fossa			X
50662	[X] Pain localized to other parts of lower abdomen			X
3869	Psychogenic dyspepsia		X	
2234	[D] Recurrent acute abdominal pain	X	X	X
15201	[D] Renal colic, unspecified			X
1181	Right iliac fossa pain			X
19360	[D] Right lower quadrant pain	X	X	X
7726	[D] Right upper quadrant pain	X		X
9695	Right upper quadrant pain	X		X
51337	Shoulder pain from abdomen			X
5960	Site of abdominal pain	X	X	X
25118	Site of GIT pain			X
29922	Site of GIT pain NOS		X	X
6357	Subcostal pain			X
20475	[D] Suprapubic pain			X
7300	Suprapubic pain			X
70357	[D] Tympanites (abdominal)	X		
14989	Type of GIT pain			X
6395	Type of GIT pain - symptom			X
44484	Type of GIT pain NOS		X	X
4771	[D] Umbilical pain	X		X
43233	Undiagnosed dyspepsia		X	
8436	[D] Upper abdominal pain	X		X
3978	Upper abdominal pain	X		X

Table 15.6 Symptom thesaurus for visible haematuria

Medcode	Description
507	Haematuria
6030	Haematuria - symptom
6234	Blood in urine - symptom
6659	Blood in urine - haematuria
6901	Clot haematuria
7164	Recurrent benign haematuria syndrome
7232	Frank haematuria
9651	Painless haematuria
13913	Blood in urine test
13915	RBCs- red blood cells in urine
13919	Urine: trace non-haemol. blood
13929	Urine blood test = +++
13932	Urine: trace haemolysed blood
13934	Urine blood test = +
17060	Recurrent and persistent haematuria
19361	Traumatic haematuria
19792	Urine blood test = ++
20357	Painful haematuria
29463	Urine microscopy:RBC's present
35555	Urine: red - blood
44541	Recurrent and persistent haematuria, dense deposit disease
47228	Essential haematuria
60856	Recur+persist haematuria difus crescentic glomerulonephritis
61317	Recur+persist haematuria difus membranous glomerulonephritis
68364	Recur+persist haematuria, focal+segmental glomerular lesions
95546	Recurrent+persistnt haematuria minor glomerular abnormality

Table 15.7 Symptom thesaurus for jaundice

Medcode	Description
355	[D]Jaundice (not of newborn)
2612	[D]Jaundice
3121	Obstructive jaundice NOS
5996	O/E - jaundiced
6000	Jaundice - symptom
18019	Yellow - symptom
18574	[D]Icterus NOS
25418	Yellow/jaundiced colour
29488	O/E - jaundiced colour
38877	[D]Jaundice (not of newborn) NOS

15.5 Appendix 5: Derivation of positive predictive value using Bayes' theorem

The positive predictive value (PPV) estimates the probability that someone has a disease given that they have a positive test result, p (disease | positive test result). Where arithmetical calculation is inappropriate, PPV can be estimated using Bayes' Theorem. Here is the derivation:

Let D be the disease (cancer) under study and S be a positive indication of disease (e.g. positive test result or presence of a symptom). From conditional probability:

The probability of disease D, given there is a positive indication (S) of disease, is:

$$p(D|S) = p(D \text{ and } S)/p(S) \quad (1)$$

Equally, the probability of S, given the presence of disease D, is:

$$p(S|D) = p(D \text{ and } S)/p(D) \quad (2)$$

So,

$$p(D \text{ and } S) = p(D|S) \times p(S) = p(S|D) \times p(D) \quad (3)$$

Substituting for $p(D \text{ and } S)$ in (1):

$$p(D|S) = [p(S|D) \times p(D)]/p(S) \quad (4)$$

= **Bayes' theorem in probability format**

If *D*+ is 'disease positive', then the alternative is 'disease negative' (*D*−); similarly, *S*+ and *S*− are symptom-positive and symptom-negative, respectively. You can write (4) in terms of the ratio of posterior probabilities:

$$\frac{p(D+\mid S+)}{p(D-\mid S+)}=\frac{p(S+\mid D+)\times p(D+)}{p(S+\mid D-)\times p(D-)} \qquad (5)$$

As far as a disease is concerned, at any one time a patient can either have it or not have it, so $p(D+)= 1-p(D-)$. This allows us to convert the above equation as odds, where odds = probability / (1 − probability):

$$odds(D+\mid S+)=odds(D+)\frac{p(S+\mid D+)}{pS+\mid D-)} \qquad (6)$$

Odds(*D*+) represents the prior odds of having the disease and can be estimated from external sources; in the CAPER studies, from national incidence data. The second term in the right-hand side of the above equation is the positive likelihood ratio (see Section 6.12.3.4). Therefore, equation (6) can be written in words as:

Posterior odds = prior odds × positive likelihood ratio (7)

Adjustment must be made for the fact that the likelihood ratio and the prior odds are each representative of different populations; namely, the consulting and the whole populations, respectively. To produce risk estimates that are meaningful to GPs, the prior odds have to be adjusted to reflect the value in the consulting

population. The adjustment factor derives from the proportions of cases and of controls who consulted their GP during the period of the study:

Posterior odds = (prior odds × *f*) × positive likelihood ratio (8)

Where $f = \frac{x_c}{x_t} \Big/ \frac{y_c}{y_t}$, in which x_c = the number of eligible cases who consulted in the study sample, x_t = the total number of cases in the study sample, y_c = the number of controls who consulted and y_t = the number of controls overall.

The assumption made here was that any patient who has not consulted the GP in the year prior to diagnosis did not have a positive indication of disease, i.e. their test results would have been normal or they did not attend for the symptom whose PPV was being estimated.

Finally, to estimate the PPV the posterior odds are converted back to a probability, as probability = odds / (1 + odds).

15.5.1 Example for illustration

I have worked through an example to illustrate the method, using coded records of visible haematuria in the bladder cancer study. We would like to estimate the probability that a patient with visible haematuria has bladder cancer. From equation (8) above, we know that:

Posterior odds of having bladder cancer = (prior odds of bladder cancer × *f*) × positive likelihood ratio of visible haematuria for bladder cancer (9)

15.5.1.1 Estimating prior odds of bladder cancer

National incidence data from the UK give us an estimate of the prior probability of having the disease, which we can convert to prior odds using the formula:

Odds = probability/(1−probability) (10)

The original bladder cancer study used UK population and bladder cancer incidence data published by the Office for National Statistics from 2008 to obtain the prior probability.[197] In 2008 in England, of the 25,101,400 men and women aged 40 years and over, there were 8,735 new diagnoses of bladder cancer, giving a prior probability of having bladder cancer equal to 0.00034799.

Substituting this in equation (10) gives us:

Odds = 0.00034799/(1−0.00034799))

= 0.0003481

Therefore, in the national population, the prior odds of having bladder cancer is 0.0003481 – and gives us the value to use in equation (9).

15.5.1.2 Adjusting the prior odds to be representative of the consulting population

The prior odds value is derived from national population data, and is not representative of the consulting population. The adjustment factor, $f = \frac{x_c}{x_t} \Big/ \frac{y_c}{y_t}$, (see above for definitions) corrects for this.

In the bladder cancer study, all 4,915 eligible and recruited cases consulted the GP in the study period; therefore, $x_c/x_t = 1$. However, of the 24,098 controls who met the inclusion criteria, 2,086 did not consult the GP in the study period; therefore, $y_c/y_t = (23{,}804-2{,}086)/23{,}804 = 0.912$.[a]

This gives a final adjustment factor, *f*, to use in equation (9):

$f = 1/0.912$

$= 1.096$

15.5.1.3 Estimating positive likelihood ratio

The positive likelihood ratio for visible haematuria is $\frac{p(S+ \mid D+)}{p(S+ \mid D-)}$, i.e. the probability of visible haematuria in bladder cancer cases divided by the probability of visible haematuria in the controls.

Coded records identified that visible haematuria occurred in 2,591 of the 4,915 (= 52.7%) cases, and in 196 of the 21,718 (0.6%) of the controls. Assuming that patients whose records contained no codes for visible haematuria never experienced the symptom, this gives us the value of the positive likelihood ratio to use in equation (9):

[a] These 2,086 patients met the exclusion criteria, as did a further 294 controls because they either received a bladder cancer diagnosis or were matched to an excluded case, leaving a final number of 21,718 controls recruited to the study (see Table 7.1).

Positive likelihood ratio = (2,591/4,915) / (196/21,718)

= **58.4 1**

This figure is representative of CPRD patients who had consulted the GP, rather than of the CPRD population as a whole.

15.5.1.4 Estimating the posterior odds

From the above figures, we can estimate the posterior odds of bladder cancer in patients who have visible haematuria using equation (9):

Posterior odds = **0.0003481 × 1.096 × 58.41**

= 0.02228

This is converted to the PPV using the formula below:

Probability (PPV) = odds / (1 + odds)

Therefore:

PPV = 0.02228 / (1 + 0.02228)

= 0.0218, or 2.18%.

16 Appendix 6: Risk estimates for the post-1998 definition of bladder cancer (C67)

This section reports the risk estimates in symptomatic patients for bladder cancer, as defined by International Classification of Disease for Oncology category C67.

Table 16.1 The positive likelihood ratio and PPV for bladder cancer (post-1998 diagnosis) in patients aged ≥40 years presenting with clinical features associated with the disease

Feature of bladder cancer	Number of patients attending at least once:		Positive likelihood ratio (95% CI)	PPV (%) (95% CI)[ab]
	Cases n (% of *n* = 3,565 cases)	Controls, n (% of *n* = 15,850 controls)		
Abdominal pain [c]	251 (7.0)	581 (3.7)	1.92 (1.66–2.22)	0.07 (0.06–0.08)
Abdominal pain [d]	389 (10.9)	896 (5.6)	1.93 (1.72–2.16)	0.07 (0.07–0.08)
Constipation [c]	205 (5.7)	507 (3.2)	1.80 (1.53–2.10)	0.07 (0.06–0.08)
Visible haematuria [c]	1,953 (54.7)	135 (0.9)	64.26 (54.19–76.22)	2.40 (2.03–2.83)
Visible haematuria [d]	2,369 (66.4)	239 (1.5)	44.03 (38.74–50.04)	1.66 (1.46–1.88)
Dysuria [c]	321 (9.0)	159 (1.0)	8.97 (7.44–10.81)	0.34 (0.28–0.41)
Urinary tract infection [c]	623 (17.5)	507 (3.2)	5.46 (4.88–6.10)	0.21 (0.19–0.23)

Feature of bladder cancer	Number of patients attending at least once:		Positive likelihood ratio (95% CI)	PPV (%) (95% CI)[ab]
	Cases n (% of *n* = 3,565 cases)	Controls, n (% of *n* = 15,850 controls)		
Raised inflammatory markers [c]	229 (6.4)	1,248 (7.9)	0.82 (0.71–0.93)	0.03 (0.03–0.04)
High white cell count [c]	189 (5.3)	294 (1.8)	2.86 (2.39–3.42)	0.11 (0.09–0.13)
Raised creatinine [c]	229 (6.4)	590 (3.7)	1.72 (1.49–2.00)	0.07 (0.06–0.08)
Non-visible haematuria [d]	242 (6.8)	46 (0.3)	23.37 (17.09–31.96)	0.88 (0.65–1.21)

[a] PPV values are adjusted for the consulting population (see Section 6.12.3.5).

[b] PPV values estimated using Bayes' Theorem (see Section 6.12.3.5), assuming a prior odds of 0.000348 based on 2008 UK national incidence data of C67 diagnoses. See Section 15.5.1 for a worked example of how to calculate PPV using Bayes' Theorem.

[c] Estimated from coded records only.

[d] Estimated from coded plus text-only records.

Table 16.2 The positive likelihood ratio and PPV for bladder cancer (post-1998 diagnosis) in patients aged ≥60 years presenting with clinical features associated with the disease

Feature of bladder cancer	Number of patients attending at least once:		Positive likelihood ratio (95% CI)	PPV (%) (95% CI)[ab]
	Cases n (% of *n* = 3,182 cases)	Controls, n (% of *n* = 14,2702 controls)		
Abdominal pain [c]	218 (6.9)	538 (3.8)	1.82 (1.56–2.12)	0.14 (0.12–0.16)
Abdominal pain [d]	350 (11.0)	839 (5.9)	1.87 (1.66–2.11)	0.14 (0.13–0.16)
Constipation [c]	197 (6.2)	491 (3.4)	1.80 (1.53–2.11)	0.14 (0.12–0.16)
Visible haematuria [c]	1,735 (54.5)	132 (0.9)	58.95 (49.59–70.06)	4.33 (3.67–5.11)
Visible haematuria [d]	2,108 (66.2)	232 (1.6)	40.75 (35.78–46.41)	3.03 (2.67–3.44)
Dysuria [c]	272 (8.5)	152 (1.1)	8.03 (6.61–9.75)	0.61 (0.50–0.74)
Urinary tract infection [c]	567(17.8)	487 (3.4)	5.22 (4.65–5.86)	0.40 (0.35–0.45)
Raised inflammatory markers [c]	212 (6.7)	555 (3.9)	1.71 (1.47–2.00)	0.13 (0.11–0.15)
High white cell count [c]	171 (5.4)	279 (2.0)	2.75 (2.28–3.31)	0.21 (0.17–0.25)
Raised creatinine [c]	483 (15.2)	1,229 (8.6)	1.76 (1.60–1.94)	0.13 (0.12–0.15)
Non-visible haematuria [d]	226 (7.1%)	46 (0.3%)	23.03 (16.08–30.18)	1.66 (1.22–2.26)

17 Permissions

For Table 4.6 **Indications for chemical dipstick testing**

Dear Sarah Price

We hereby grant you permission to reproduce the material detailed below at no charge **in your thesis, in print and on Open Research Exeter (ORE).**subject to the following conditions:

1. If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies.
2. Suitable acknowledgment to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:
"This article was published in Publication title, Vol number, Author(s), Title of article, Page Nos, Copyright Elsevier (or appropriate Society name) (Year)."
3. Your thesis may be submitted to your institution in either print or electronic form.
4. Reproduction of this material is confined to the purpose for which permission is hereby given.
5. This permission is granted for non-exclusive world **English** rights only. For other languages please reapply separately for each one required. Permission excludes use in an electronic form other than as specified above. Should you have a specific electronic project in mind please reapply for permission.
6. Should your thesis be published commercially, please reapply for permission.

Yours sincerely

cid:image002.gif@01CFACBC.AE7A6640
Jennifer Jones
Rights Associate

Elsevier Limited, a company registered in England and Wales with company number 1982084, whose registered office is The Boulevard, Langford Lane, Kidlington, Oxford, OX5 1GB, United Kingdom.

Please select the type of publication:

Book

Book - Title:

Macleod's Clinical Examination

Book - ISBN:

978-0443068485

Book - Author(s):

Graham Douglas, Fiona Nicol, Colin Robertson

Book - Year:

2009

Book - Pages from:

1

Book - Pages to:

476

Book - Chapter Num:

9

Book - Chapter Title:

The renal system

I would like to use (please select one of the following options):

Table(s)

If using figures/tables or illustrations please specify the quantity:

1 table

Are you the author of the material?:

No

If not, is the author involved with your project:

No

In what format will you use the material?:

Print

Will you be translating the material?:

No

Information about your proposed use:

thesis

Proposed use text:

All thesis are held in an open-access repository called Open Research Exeter (ORE). Research can be viewed and downloaded freely by anyone, anywhere: researchers, students, industry, business and the wider public.

For Table 4.7 Causes of isolated non-visible haematuria (causes are listed in order of descending frequency of presentation, according to available data)

Reuse of Content within a Thesis or Dissertation

Content (full-text or portions thereof) may be used in print and electronic versions of your dissertation or thesis without formal permission from the Massachusetts Medical Society, Publisher of the New England Journal of Medicine.

The following credit line must be printed along with the copyrighted material:

"Reproduced with permission from (scientific reference citation), Copyright Massachusetts Medical Society.

Figure 4.7 Some important skin areas involved in referred visceral pain (Figure 5-70 reproduced with permission from Richard S. Snell, Clinical Anatomy for Medical Students, 5th edition, © Richard S. Snell MD PhD, 199548)

Dear Sarah,

Thank you for contacting Wolters Kluwer with your permission request. We do allow Figure 5-70 from Snell: Clinical Anatomy for Medical Students 5e, 9780316801355 to be borrowed and will be able to grant your request. I am attaching a copy of our Terms and Conditions. Please consider those, and this email, your official grant of permission.

Health Permissions Team

Health Learning, Research & Practice

Wolters Kluwer

healthpermissions@wolterskluwer.com

18 Glossary of terms

Haematuria, visible The abnormal presence of red blood cells in urine in sufficient quantities to be readily recognised by the patient. Synonyms: frank, gross, macroscopic

Haematuria, non-visible The abnormal presence of red blood cells in urine in an amount so small that it requires detection by urine dipstick testing or microscopy. Synonyms: invisible, dipstick-positive, microscopic

Medcode The numeric form of the alphanumeric Read code, generated exclusively by the CPRD to facilitate data management, storage and manipulation in software packages such as Stata (*see* Read codes)

Noun phrase A word or group of words containing a noun and functioning in a sentence as subject, object, or prepositional object.

Read codes A thesaurus of clinical events, each with a unique alphanumeric code, that has been used in the NHS since 1985. There are two versions: version 2 (v2) and version 3 (CTV3 or v3). Read codes provide a standard vocabulary for clinicians to record patient findings and procedures in health and social care IT systems across primary and secondary care (e.g. General Practice surgeries and pathology reporting of results). For more information see http://systems.hscic.gov.uk/data/uktc/readcodes. The CPRD receive data from contributing practices in the form of alphanumeric Read codes, and map them

1:1 to numeric medcodes, purely for ease of data management, storage and manipulation in software packages such as Stata (*see* Medcodes)

Syntax The arrangement of words and phrases to create well-formed sentences in a language.

Verb phrase A verb with another word or words indicating tense, mood, or person.

19 Bibliography

1. Williams T, Tjeerd van S, Puri S, Eaton S. Recent advances in the utility and use of the General Practice Research Database as an example of a UK primary care data resource. *Therapeutic Advances in Drug Safety* 2012;**3**(2):89-99.

2. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data resource profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015;**44**(3):827-36.

3. The General Practice Research Database. GPRD recording guidelines for Vision users. London: Crown Publishing, 2004.

4. Shephard E, Stapley S, Neal RD, Rose P, Walter F, Hamilton W. Clinical features of bladder cancer in primary care. *Br J Gen Pract* 2012;**62**:e598-e604.

5. Stapley S, Peters TJ, Neal RD, Rose PW, Walter FM, Hamilton W. The risk of pancreatic cancer in symptomatic patients in primary care: A large case-control study using electronic records. *Br J Cancer* 2012;**106**(12):1940-4.

6. Hippisley-Cox J, Coupland C. Identifying patients with suspected renal tract cancer in primary care: Derivation and validation of an algorithm. *Br J Gen Pract* 2012;**62**(597):e251-60.

7. Hippisley-Cox J, Coupland C. Identifying patients with suspected pancreatic cancer in primary care: Derivation and validation of an algorithm. *Br J Gen Pract* 2012;**62**(594):e38-45.

8. Price SJ, Shephard EA, Stapley SA, Barraclough K, Hamilton WT. Non-visible versus visible haematuria and bladder cancer risk: A study of electronic records in primary care. *Br J Gen Pract* 2014;**64**(626):e584-9.

9. National Institute for Health and Care Excellence. Suspected cancer: Recognition and referral [NG12]. 2015 23rd June, 2015. http://www.nice.org.uk/guidance/NG12.

10. Garcia Rodriguez LA, Perez Gutthann S. Use of the UK General Practice Research Database for pharmacoepidemiology. *Br J Clin Pharmacol* 1998;**45**(5):419-25.

11. Hamilton WT, Round AP, Sharp D, Peters TJ. The quality of record keeping in primary care: A comparison of computerised, paper and hybrid systems. *Br J Gen Pract* 2003;**53**(497):929-33; discussion 33.

12. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: A perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association : JAMIA* 2011;**18**(2):181-86.

13. Fernando B, Kalra D, Morrison Z, Byrne E, Sheikh A. Benefits and risks of structuring and/or coding the presenting patient history in the electronic health record: Systematic review. *BMJ Quality & Safety* 2012;**21**(4):337-46.

14. Medicines and Healthcare Products Regulatory Agency T. The General Practice Research Database. London: The Medicines and Healthcare Products Regulatory Agency (Undated).

15. Nicholson A, Tate AR, Koeling R, Cassell JA. What does validation of cases in electronic record databases mean? The potential contribution of free text. *Pharmacoepidemiol Drug Saf* 2011;**20**(3):321-4.

16. Shephard E, Stapley S, Hamilton W. The use of electronic databases in primary care research. *Fam Pract* 2011;**28**(4):352-4.

17. Lewis JD, Schinnar R, Bilker WB, Wang X, Strom BL. Validation studies of The Health Improvement Network (THIN) database for pharmacoepidemiology research. *Pharmacoepidemiol Drug Saf* 2007;**16**(4):393-401.

18. Hippisley-Cox J, Stables D, Pringle M. QResearch: A new general practice database for research. *Inform Prim Care* 2004;**12**(1):49-50.

19. Hippisley-Cox J, Coupland C. Development and validation of risk prediction equations to estimate future risk of blindness and lower limb amputation in patients with diabetes: Cohort study. *BMJ* 2015;**351**:h5441.

20. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: A systematic review. *Br J Clin Pharmacol* 2010;**69**(1):4-14.

21. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: A systematic review. *The British Journal Of General Practice* 2010;**60**(572):e128-e36.

22. Wurst KE, Ephross SA, Loehr J, Clark DW, Guess HA. The utility of the General Practice Research Database to examine selected congenital heart defects: A validation study. *Pharmacoepidemiol Drug Saf* 2007;**16**(8):867-77.

23. Devine S, West S, Andrews E, Tennis P, Hammad TA, Eaton S, et al. The identification of pregnancies within the General Practice Research Database. *Pharmacoepidemiol Drug Saf* 2010;**19**(1):45-50.

24. Charlton RA, Weil JG, Cunnington MC, de Vries CS. Identifying major congenital malformations in the UK General Practice Research Database (GPRD): A study reporting on the sensitivity and added value of photocopied medical records and free text in the GPRD. *Drug Saf* 2010;**33**(9):741-50.

25. Boggon R, van Staa TP, Chapman M, Gallagher AM, Hammad TA, Richards MA. Cancer recording and mortality in the General Practice Research Database and linked cancer registries. *Pharmacoepidemiol Drug Saf* 2013;**22**(2):168-75.

26. Close H, Reilly J, Mason JM, Kripalani M, Wilson D, Main J, et al. Renal failure in lithium-treated bipolar disorder: A retrospective cohort study. *PLoS One* 2014;**9**(3):e90169-e69.

27. Thomas KH, Davies N, Metcalfe C, Windmeijer F, Martin RM, Gunnell D. Validation of suicide and self-harm records in the Clinical Practice Research Datalink. *Br J Clin Pharmacol* 2013;**76**(1):145-57.

28. Shah AD, Martinez C, Hemingway H. The freetext matching algorithm: A computer program to extract diagnoses and causes of death from unstructured text in electronic health records. *BMC Med Inform Decis Mak* 2012;**12**:88.

29. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: A review of recent research. *Yearb Med Inform* 2008:128-44.

30. Tate AR, Martin AG, Ali A, Cassell JA. Using free text information to explore how and when GPs code a diagnosis of ovarian cancer: An observational study using primary care records of patients with ovarian cancer. *BMJ Open* 2011;**1**(1):e000025.

31. Nicholson A, Ford E, Davies KA, Smith HE, Rait G, Tate AR, et al. Optimising use of electronic health records to describe the presentation of rheumatoid arthritis in primary care: A strategy for developing code lists. *PLoS One* 2013;**8**(2):e54878.

32. Ford E, Nicholson A, Koeling R, Tate AR, Carroll J, Axelrod L, et al. Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: What information is hidden in free text? *BMC Med Res Methodol* 2013;**13**(1):105.

33. Koeling R, Tate AR, Carroll JA. Automatically estimating the incidence of symptoms recorded in GP free text notes. First International Workshop on Managing Interoperability and Complexity in Health Systems. Glasgow, Scotland, 2011.

34. Cancer Research UK. Cancerstats. . Secondary Cancerstats. 2012. http://www.cancerresearchuk.org.

35. Abdel-Rahman M, Stockton D, Rachet B, Hakulinen T, Coleman MP. What if cancer survival in Britain were the same as in Europe: How many deaths are avoidable? *Br J Cancer* 2009;**101 Suppl 2**:S115-24.

36. HM Government. The NHS Cancer Plan: A plan for investment, a plan for reform. London: Department of Health, 2000.

37. HM Government. The national cancer strategy. London: The Department of Health, 2011.

38. Hamilton W, Green T, Martins T, Elliott K, Rubin G, Macleod U. Evaluation of risk assessment tools for suspected cancer in general practice: A cohort study. *Br J Gen Pract* 2013;**63**(606):e30-6.

39. Usher-Smith J, Emery J, Hamilton W, Griffin SJ, Walter FM. Risk prediction tools for cancer in primary care. *Br J Cancer* 2015;**113**(12):1645-50.

40. Shephard E, Neal R, Rose P, Walter F, Hamilton WT. Clinical features of kidney cancer in primary care: A case-control study using primary care records. *Br J Gen Pract* 2013;**63**(609):e250-5.

41. Stapley S, Peters TJ, Neal RD, Rose PW, Walter FM, Hamilton W. The risk of oesophago-gastric cancer in symptomatic patients in primary care: A large case-control study using electronic records. *Br J Cancer* 2013;**108**(1):25-31.

42. Walker S, Hyde C, Hamilton W. Risk of breast cancer in symptomatic women in primary care: A case-control study using electronic records. *Br J Gen Pract* 2014;**64**(629):e788-93.

43. Walker S, Hyde C, Hamilton W. Risk of uterine cancer in symptomatic women in primary care: Case-control study using electronic records. *Br J Gen Pract* 2013;**63**(614):e643-8.

44. Shephard EA, Neal RD, Rose P, Walter FM, Litt EJ, Hamilton WT. Quantifying the risk of multiple myeloma from symptoms reported in primary care patients: A large case-control study using electronic records. *Br J Gen Pract* 2015;**65**(631):e106-13.

45. Shephard EA, Neal RD, Rose PW, Walter FM, Hamilton WT. Quantifying the risk of non-hodgkin lymphoma in symptomatic primary care patients aged >/=40

years: A large case-control study using electronic records. *Br J Gen Pract* 2015;**65**(634):e281-8.

46. Shephard EA, Neal RD, Rose PW, Walter FM, Hamilton WT. Quantifying the risk of hodgkin lymphoma in symptomatic primary care patients aged >/=40 years: A case-control study using electronic records. *Br J Gen Pract* 2015;**65**(634):e289-94.

47. Health and Social Care Information Centre. Quality and Outcomes Framework. Secondary Quality and Outcomes Framework. 2015. http://qof.hscic.gov.uk/index.asp.

48. Snell RS. *Clinical anatomy for medical students*. 5th ed. Boston: Little, Brown and Company, 1995.

49. Davidson S. *Davidson's principles and practice of medicine*. 17th ed. New York: Churchill Livingstone, 1995.

50. Parkin DM, Boyd L, Walker LC. 16. The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010. *Br J Cancer* 2011;**105 Suppl 2**:S77-81.

51. Hidalgo M. Pancreatic cancer. *N Engl J Med* 2010;**362**(17):1605-17.

52. Jacobs EJ, Chanock SJ, Fuchs CS, Lacroix A, McWilliams RR, Steplowski E, et al. Family history of cancer and risk of pancreatic cancer: A pooled analysis from

the pancreatic cancer cohort consortium (PANSCAN). *Int J Cancer* 2010;**127**(6):1421-8.

53. Cancer Research UK. Pancreatic cancer statistics. Secondary Pancreatic cancer statistics. 2011. http://www.cancerresearchuk.org/cancer-info/cancerstats/types/pancreas/incidence/#age.

54. Vincent A, Herman J, Schulick R, Hruban RH, Goggins M. Pancreatic cancer. *Lancet* 2011;**378**(9791):607-20.

55. Schmidt-Hansen M, Berendse S, Hamilton W. Symptoms of pancreatic cancer in primary care: A systematic review. *Pancreas* 2015.

56. National Cancer Intelligence Network. Routes to diagnosis 2006-2010 workbook. London: NCIN, 2013.

57. National Institute for Health and Clinical Excellence (NICE). *Referral guidelines for suspected cancer* London: NICE, 2005.

58. Lyratzopoulos G, Neal RD, Barbiere JM, Rubin GP, Abel GA. 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**(4):353-65.

59. Bilimoria KY, Bentrem DJ, Ko CY, Ritchey J, Stewart AK, Winchester DP, et al. Validation of the 6th edition AJCC Pancreatic Cancer Staging System: Report from the National Cancer Database. *Cancer* 2007;**110**(4):738-44.

60. Office for National Statistics. Mortality statistics: Deaths registered in England and Wales (series DR), 2013. In: Office for National Statistics, ed. London Office for National Statistics, 2013.

61. Cancer Research UK. Bladder cancer statistics. Secondary Bladder cancer statistics 2014. http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bladder-cancer#heading-Zero.

62. Ellis L, Woods LM, Esteve J, Eloranta S, Coleman MP, Rachet B. Cancer incidence, survival and mortality: Explaining the concepts. *Int J Cancer* 2014;**135**(8):1774-82.

63. Lifestyle Statistics. Statistics on smoking: England, 2013. In: Health and Social Care Information Centre, ed. London: Health and Social Care Information Centre, 2013.

64. International Agency for Research on Cancer. IARC monographs on the evaluation of carcinogenic risk to humans. 2015. http://monographs.iarc.fr/ENG/Classification/index.php.

65. Schmidt-Hansen M, Berendse S, Hamilton W. The association between symptoms and bladder or renal tract cancer in primary care: A systematic review. *Br J Gen Pract* 2015;**65**(640):e769-75.

66. Bruyninckx R, Buntinx F, Aertgeerts B, Van Casteren V. The diagnostic value of macroscopic haematuria for the diagnosis of urological cancer in general practice. *Br J Gen Pract* 2003;**53**(486):31-5.

67. Collins GS, Altman DG. Identifying patients with undetected renal tract cancer in primary care: An independent and external validation of QCancer(r) (renal) prediction model. *Cancer Epidemiol* 2013;**37**(2):115-20.

68. Friedlander DF, Resnick MJ, You C, Bassett J, Yarlagadda V, Penson DF, et al. Variation in the intensity of hematuria evaluation: A target for primary care quality improvement. *Am J Med* 2014;**127**(7):633-40.e11.

69. Jones R, Latinovic R, Charlton J, Gulliford MC. Alarm symptoms in early diagnosis of cancer in primary care: Cohort study using General Practice Research Database. *BMJ* 2007;**334**(7602):1040.

70. Nicholson BD, McGrath JS, Hamilton W. Bladder cancer in women. *BMJ* 2014;**348**:g2171.

71. Guyton AC, Hall JE. *Textbook of medical physiology*. 11th ed. Philadelphia: Saunders, 2005.

72. Rodgers M, Nixon J, Hempel S, Aho T, Kelly J, Neal D, et al. Diagnostic tests and algorithms used in the investigation of haematuria: Systematic reviews and economic evaluation. *Health Technol Assess* 2006;**10**(18):iii-iv, xi-259.

73. Kelly JD, Fawcett DP, Goldberg LC. Assessment and management of non-visible haematuria in primary care. *BMJ* 2009;**338**:a3021.

74. National Collaborating Centre for Chronic Conditions. *National Institute for Health and Clinical Excellence: Guidance. Chronic kidney disease: National clinical guideline for early identification and management in adults in primary and secondary care.* London: Royal College of Physicians (UK), 2008.

75. Summerton N, Mann S, Rigby AS, Ashley J, Palmer S, Hetherington JW. Patients with new onset haematuria: Assessing the discriminant value of clinical information in relation to urological malignancies. *Br J Gen Pract* 2002;**52**(477):284-9.

76. Froom P, Ribak J, Benbassat J. Significance of microhaematuria in young adults. *Br Med J (Clin Res Ed)* 1984;**288**(6410):20-2.

77. Cohen RA, Brown RS. Clinical practice. Microscopic hematuria. *N Engl J Med* 2003;**348**(23):2330-8.

78. Hiatt RA, Ordonez JD. Dipstick urinalysis screening, asymptomatic microhematuria, and subsequent urological cancers in a population-based sample. *Cancer Epidemiol Biomarkers Prev* 1994;**3**(5):439-43.

79. Ritchie CD, Bevan EA, Collier SJ. Importance of occult haematuria found at screening. *Br Med J (Clin Res Ed)* 1986;**292**(6521):681-3.

80. Britton JP, Dowell AC, Whelan P. Dipstick haematuria and bladder cancer in men over 60: Results of a community study. *BMJ* 1989;**299**(6706):1010-2.

81. Jones GR, Newhouse I. Sport-related hematuria: A review. *Clin J Sport Med* 1997;**7**(2):119-25.

82. Macleod J. *Macleod's clinical examination*. 12th ed. Edinburgh: Churchill Livingstone, 2009.

83. Stryer L. *Biochemistry*. 3rd ed. New York: Freeman, 1988.

84. Grossfeld GD, Litwin MS, Wolf JS, Hricak H, Shuler CL, Agerter DC, et al. Evaluation of asymptomatic microscopic hematuria in adults: The American Urological Association best practice policy--part I: Definition, detection, prevalence, and etiology. *Urology* 2001;**57**(4):599-603.

85. Longmore M, Wilkinson IB, Baldwin A, Wallin E. *Oxford handbook of clinical medicine*. 9th ed. Oxford: Oxford University Press, 2014.

86. Rao PK, Jones JS. How to evaluate 'dipstick hematuria': What to do before you refer. *Cleve Clin J Med* 2008;**75**(3):227-33.

87. Grossfeld GD, Litwin MS, Wolf JS, Jr., Hricak H, Shuler CL, Agerter DC, et al. Evaluation of asymptomatic microscopic hematuria in adults: The American Urological Association best practice policy--part II: Patient evaluation, cytology, voided markers, imaging, cystoscopy, nephrology evaluation, and follow-up. *Urology* 2001;**57**(4):604-10.

88. Grossfeld GD, Carroll PR. Evaluation of asymptomatic microscopic hematuria. *Urol Clin North Am* 1998;**25**(4):661-76.

89. Sadow CA, Silverman SG, O'Leary MP, Signorovitch JE. Bladder cancer detection with CT urography in an academic medical center. *Radiology* 2008;**249**(1):195-202.

90. Beckingham IJ, Ryder SD. Abc of diseases of liver, pancreas, and biliary system. Investigation of liver and biliary disease. *BMJ* 2001;**322**(7277):33-6.

91. Browse N. *An introduction to the symptoms and signs of surgical disease*. 3rd ed: Taylor & Francis, 1997.

92. Taylor A, Stapley S, Hamilton W. Jaundice in primary care: A cohort study of adults aged >45 years using electronic medical records. *Fam Pract* 2012;**29**(4):416-20.

93. Roche SP, Kobos R. Jaundice in the adult patient. *Am Fam Physician* 2004;**69**(2):299-304.

94. Hutton P, Cooper G, James FM, Butterworth JF. *Fundamental principles and practice of anaesthesia*: Taylor & Francis, 2002.

95. Pinnock C, Lin T, Smith T, Jones R. *Fundamentals of anaesthesia*. London: Greenwich Medical Media Ltd, 1999.

96. Breivik H, Borchgrevink PC, Allen SM, Rosseland LA, Romundstad L, Hals EK, et al. Assessment of pain. *Br J Anaesth* 2008;**101**(1):17-24.

97. General Medical Council. Good medical practice. Manchester: General Medical Council, 2013.

98. Pelaccia T, Tardif J, Triby E, Charlin B. An analysis of clinical reasoning through a recent and comprehensive approach: The dual-process theory. *Med Educ Online* 2011;**16**.

99. Mann RD, Andrews EB. *Pharmacovigilance.* New York: John Wiley & Sons, 2007.

100. Breslow NE, Day NE. Statistical methods in cancer research. Volume I - the analysis of case-control studies. *IARC Sci Publ* 1980(32):5-338.

101. Hippisley-Cox J, Vinogradova Y. Trends in consultation rates in general practice 1995 to 2008: Analysis of the QResearch® database. London, 2009.

102. Wang Y, Hunt K, Nazareth I, Freemantle N, Petersen I. Do men consult less than women? An analysis of routinely collected UK general practice data. *BMJ Open* 2013;**3**(8):e003320.

103. Wang Y, Freemantle N, Nazareth I, Hunt K. Gender differences in survival and the use of primary care prior to diagnosis of three cancers: An analysis of routinely collected UK general practice data. *PLoS One* 2014;**9**(7):e101562.

104. Hamilton W. Towards earlier diagnosis of cancer in primary care: A population-based case-control study of colorectal, lung and prostate cancer. MD Thesis, University of Bristol, 2005.

105. McCormick A, Fleming D, Charlton J. *Morbidity statistics from general practice: Fourth national study 1991-1992*. London: Her Majesty's Stationery Office, 1995.

106. Norman G, Barraclough K, Dolovich L, Price D. Iterative diagnosis. *BMJ* 2009;**339**.

107. Mann R, Williams J. Standards in medical record keeping. *Clin Med (Lond)* 2003;**3**(4):329-32.

108. Heneghan C, Glasziou P, Thompson M, Rose P, Balla J, Lasserson D, et al. Diagnostic strategies used in primary care. *BMJ* 2009;**338**:b946.

109. Elstein AS, Shulman LS, Sprafka SA. Medical problem-solving. *J Med Educ* 1981;**56**(1):75-6.

110. Marewski JN, Gigerenzer G. Heuristic decision making in medicine. *Dialogues Clin Neurosci* 2012;**14**(1):77-89.

111. Fraser RC. *Clinical method: A general practice approach*. Second Revised ed. Oxford: Butterworth Heinemann, 1997.

112. Horn LR. *A natural history of negation.* Chicago: University of Chicago Press, 1989.

113. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess* 2007;**11**(50):iii, ix-51.

114. Kirkwood BR, Sterne, Jonathan A. C. . *Essential medical statistics*. 2nd ed. Oxford: Blackwell Science, 2003.

115. Aronow DB, Fangfang F, Croft WB. Ad hoc classification of radiology reports. *J Am Med Inform Assoc* 1999;**6**(5):393-411.

116. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 2001;**34**(5):301-10.

117. Mutalik PG, Deshpande A, Nadkarni PM. Use of general-purpose negation detection to augment concept indexing of medical documents: A quantitative study using the UMLS. *J Am Med Inform Assoc* 2001;**8**(6):598-609.

118. Wang Z, Shah AD, Tate AR, Denaxas S, Shawe-Taylor J, Hemingway H. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. *PLoS One* 2012;**7**(1):e30412.

119. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990;**43**(6):543-9.

120. Altman DG. *Practical statistics for medical research.* London: Chapman & Hall, 1991.

121. Landis J, Koch G. The measurement of interobserver agreement for categorical data. *Biometrics* 1977;**33**:159-74.

122. McGinn T, Wyer PC, Newman TB, Keitz S, Leipzig R, For GG. Tips for learners of evidence-based medicine: 3. Measures of observer variability (kappa statistic). *CMAJ* 2004;**171**(11):1369-73.

123. WONCA International Classification Committee. *International classification of primary care ICPC-2-r*. Revised Second Edition. Oxford: Oxford University Press, 2005.

124. D'Agostino RB, Belanger A, D'Agostino J, Ralph B. . A suggestion for using powerful and informative tests of normality. *The American Statistician* 1990;**44**(4):316-21.

125. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based medicine: How to practice and teach EBM*. 2nd ed. London: Churchill Livingstone, 2000.

126. Holtedahl KA. A method of calculating diagnostic indexes for possible cancer symptoms in general practice. *Allgemein Medizin* 1990;**19**:74-79.

127. Shephard EA, Neal RD, Rose PW, Walter FM, Hamilton W. Symptoms of adult chronic and acute leukaemia before diagnosis: Large primary care case-control studies using electronic records. *Br J Gen Pract* 2016;**66**(644):e182-8.

128. Stapley S, Peters TJ, Sharp D, Hamilton W. The mortality of colorectal cancer in relation to the initial symptom at presentation to primary care and to the

duration of symptoms: A cohort study using medical records. *Br J Cancer* 2006;**95**(10):1321-5.

129. Hamilton W, Hajioff S, Graham J, Schmidt-Hansen M. Suspected cancer (part 2-adults): Reference tables from updated NICE guidance. *BMJ* 2015;**350**:h3044.

130. Hosmer DW, Lemeshow S. *Applied logistic regression*. New York: John Wiley & sons, 1989.

131. Long JS, Freese J. *Regression models for categorical dependent variables using stata*. 3rd ed. College Station, Texas: Stata Press, 2014.

132. Dregan A, Moller H, Murray-Thomas T, Gulliford MC. Validity of cancer diagnosis in a primary care database compared with linked cancer registrations in England. Population-based cohort study. *Cancer Epidemiol* 2012;**36**(5):425-9.

133. Bhaskaran K, Forbes HJ, Douglas I, Leon DA, Smeeth L. Representativeness and optimal use of body mass index (BMI) in the UK Clinical Practice Research Datalink (CPRD). *BMJ Open* 2013;**3**(9):e003389.

134. Booth HP, Prevost AT, Gulliford MC. Validity of smoking prevalence estimates from primary care electronic health records compared with national population survey data for England, 2007 to 2011. *Pharmacoepidemiol Drug Saf* 2013;**22**(12):1357-61.

135. Dommett RM, Redaniel MT, Stevens MC, Hamilton W, Martin RM. Features of childhood cancer in primary care: A population-based nested case-control study. *Br J Cancer* 2012;**106**(5):982-7.

136. Delgado-Rodriguez M, Llorca J. Bias. *J Epidemiol Community Health* 2004;**58**(8):635-41.

137. Altman DG, Bland JM. Missing data. *BMJ* 2007;**334**(7590):424.

138. General Practitioners Committee, Royal College of General Practitioners, Department of Health. Good practice guidelines for general practice electronic patient records. London: Department of Health, 2005.

139. Department of Health, Royal College of General Practitioners, British Medical Association. The good practice guidelines for GP electronic patient records. London: Department of Health, 2011.

140. The Medical Defence Union Limited. Report and accounts 2014. London: MDU, 2014.

141. HM Government. Criminal justice and courts act 2015. Norwich: The Stationery Office, 2015.

142. HM Government. The health care and associated professions (indemnity arrangements) order 2014. Norwich: The Stationery Office, 2014.

143. Cancer Research UK. Statistics and outlook for pancreatic cancer. Secondary Statistics and outlook for pancreatic cancer. 2015. http://www.cancerresearchuk.org/about-cancer/type/pancreatic-cancer/treatment/statistics-and-outlook-for-pancreatic-cancer.

144. Buntinx F, Mant D, Van den Bruel A, Donner-Banzhof N, Dinant GJ. Dealing with low-incidence serious diseases in general practice. *Br J Gen Pract* 2011;**61**(582):43-6.

145. Stolper E, van Bokhoven M, Houben P, Van Royen P, van de Wiel M, van der Weijden T, et al. The diagnostic role of gut feelings in general practice. A focus group study of the concept and its determinants. *BMC Fam Pract* 2009;**10**:17.

146. Stolper E, Van de Wiel M, Van Royen P, Van Bokhoven M, Van der Weijden T, Dinant GJ. Gut feelings as a third track in general practitioners' diagnostic reasoning. *J Gen Intern Med* 2011;**26**(2):197-203.

147. Stolper E, Van Royen P, Van de Wiel M, Van Bokhoven M, Houben P, Van der Weijden T, et al. Consensus on gut feelings in general practice. *BMC Fam Pract* 2009;**10**:66.

148. Elstein AS, Schwartz A. Clinical problem solving and diagnostic decision making: Selective review of the cognitive literature. *BMJ* 2002;**324**(7339):729-32.

149. Djulbegovic B, Hozo I, Beckstead J, Tsalatsanis A, Pauker SG. Dual processing model of medical decision-making. *BMC Med Inform Decis Mak* 2012;**12**:94.

150. Kahneman D. *Thinking, fast and slow*. London: Penguin, 2011.

151. Van den Bruel A, Aertgeerts B, Bruyninckx R, Aerts M, Buntinx F. Signs and symptoms for diagnosis of serious infections in children: A prospective study in primary care. *Br J Gen Pract* 2007;**57**(540):538-46.

152. Elstein AS. Thinking about diagnostic thinking: A 30-year perspective. *Adv Health Sci Educ Theory Pract* 2009;**14 Suppl 1**:7-18.

153. Norman GR, Eva KW. Diagnostic error and clinical reasoning. *Med Educ* 2010;**44**(1):94-100.

154. Marcum JA. An integrated model of clinical reasoning: Dual-process theory of cognition and metacognition. *J Eval Clin Pract* 2012;**18**(5):954-61.

155. Health and Social Care Information Centre. National quality and outcomes framework statistics for england 2005/05. London: Health and Social Care Information Centre, 2004.

156. Hamilton W, Peters TJ, Bankhead C, Sharp D. Risk of ovarian cancer in women with symptoms in primary care: Population based case-control study. *BMJ* 2009;**339**:b2998.

157. Hayward RA, Chen Y, Croft P, Jordan KP. Presentation of respiratory symptoms prior to diagnosis in general practice: A case-control study examining free text and morbidity codes. *BMJ Open* 2015;**5**(6):e007355.

158. Lyratzopoulos G, Abel GA, McPhail S, Neal RD, Rubin GP. Gender inequalities in the promptness of diagnosis of bladder and renal cancer after symptomatic presentation: Evidence from secondary analysis of an English primary care audit survey. *BMJ Open* 2013;**3**(6).

159. Bossuyt PM, Davenport C, Deeks J, Hyde C, Leeflang M, Scholten R. Interpreting results and drawing conclusions. In: Deeks J, Bossuyt PM, Gatsonis CA, eds. *Cochrane handbook for systematic reviews of diagnostic test accuracy*. London: Cochrane Collaboration, 2013.

160. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCAP)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009;**42**(2):377-81.

161. Edwards TJ, Dickinson AJ, Natale S, Gosling J, McGrath JS. A prospective analysis of the diagnostic yield resulting from the attendance of 4020 patients at a protocol-driven haematuria clinic. *BJU Int* 2006;**97**(2):301-5; discussion 05.

162. Shah AD, Martinez C. An algorithm to derive a numerical daily dose from unstructured text dosage instructions. *Pharmacoepidemiol Drug Saf* 2006;**15**(3):161-6.

163. Black C, Jick H. Etiology and frequency of rhabdomyolysis. *Pharmacotherapy* 2002;**22**(12):1524-6.

164. Tate AR, Martin AGR, Murray-Thomas T, Anderson SR, Cassell JA. Determining the date of diagnosis--is it a simple matter? The impact of different approaches to dating diagnosis on estimates of delayed care for ovarian cancer in UK primary care. *BMC Med Res Methodol* 2009;**9**:42-42.

165. Bhavnani V, Fisher B, Winfield M, Seed P. How patients use access to their electronic GP record--a quantitative study. *Fam Pract* 2011;**28**(2):188-94.

166. Byrne E, Fernando B, Kalra D, Sheikh A. The benefits and risks of structuring and coding of patient histories in the electronic clinical record: Protocol for a systematic review. *Inform Prim Care* 2010;**18**(3):197-203.

167. Flynn RWV, Macdonald TM, Schembri N, Murray GD, Doney ASF. Automated data capture from free-text radiology reports to enhance accuracy of hospital inpatient stroke codes. *Pharmacoepidemiol Drug Saf* 2010;**19**(8):843-47.

168. Hunter J, Freer Y, Gatt A, Reiter E, Sripada S, Sykes C, et al. Bt-Nurse: Computer generation of natural language shift summaries from complex heterogeneous medical data. *Journal Of The American Medical Informatics Association: JAMIA* 2011;**18**(5):621-24.

169. Ruigómez A, Johansson S, Wernersson B, Fernández Cantero O, García Rodríguez LA. Gastroesophageal reflux disease in primary care: Using changes in proton pump inhibitor therapy as an indicator of partial response. *Scand J Gastroenterol* 2012;**47**(7):751-61.

170. Bannister CA, Poole CD, Jenkins-Jones S, Morgan CL, Elwyn G, Spasic I, et al. External validation of the ukpds risk engine in incident type 2 diabetes: A need for new type 2 diabetes-specific risk equations. *Diabetes Care* 2014;**37**(2):537-45.

171. Hammad TA, Margulis AV, Ding Y, Strazzeri MM, Epperly H. Determining the predictive value of read codes to identify congenital cardiac malformations in the UK Clinical Practice Research Datalink. *Pharmacoepidemiol Drug Saf* 2013;**22**(11):1233-8.

172. Hippisley-Cox J, Coupland C. Predicting risk of emergency admission to hospital using primary care data: Derivation and validation of QAdmissions score. *BMJ Open* 2013;**3**(8):e003482-e82.

173. Pouwels S, Bazelier MT, de Boer A, Weber WEJ, Neef CK, Cooper C, et al. Five-year fracture risk estimation in patients with Parkinson's disease. *Bone* 2013;**56**(2):266-70.

174. Reeves D, Springate DA, Ashcroft DM, Ryan R, Doran T, Morris R, et al. Can analyses of electronic patient records be independently and externally validated? The effect of statins on the mortality of patients with ischaemic heart disease: A cohort study with nested case-control analysis. *BMJ Open* 2014;**4**(4):e004952-e52.

175. Cornish RP, Henderson J, Boyd AW, Granell R, Van Staa T, Macleod J. Validating childhood asthma in an epidemiological study using linked electronic patient records. *BMJ Open* 2014;**4**(4):e005345-e45.

176. de Abajo FJ, Rodríguez LA, Montero D. Association between selective serotonin reuptake inhibitors and upper gastrointestinal bleeding: Population based case-control study. *BMJ (Clinical Research Ed)* 1999;**319**(7217):1106-09.

177. Derby L, Maier WC. Risk of cataract among users of intranasal corticosteroids. *The Journal Of Allergy And Clinical Immunology* 2000;**105**(5):912-16.

178. Devine S, West SL, Andrews E, Tennis P, Eaton S, Thorp J, et al. Validation of neural tube defects in the full featured--General Practice Research Database. *Pharmacoepidemiol Drug Saf* 2008;**17**(5):434-44.

179. Egbring M, Kullak-Ublick GA, Russmann S. Phynx: An open source software solution supporting data management and web-based patient-level data review for drug safety studies in the General Practice Research Database and other health care databases. *Pharmacoepidemiol Drug Saf* 2010;**19**(1):38-44.

180. Hall GC, Morant SV, Carroll D, Gabriel ZL, McQuay HJ. An observational descriptive study of the epidemiology and treatment of neuropathic pain in a UK general population. *BMC Fam Pract* 2013;**14**:28-28.

181. Hoffmann F, Andersohn F, Giersiepen K, Scharnetzky E, Garbe E. [validation of secondary data. Strengths and limitations]. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz* 2008;**51**(10):1118-26.

182. Huerta C, Castellsague J, Varas-Lorenzo C, García Rodríguez LA. Nonsteroidal anti-inflammatory drugs and risk of ARF in the general population. *American Journal Of Kidney Diseases: The Official Journal Of The National Kidney Foundation* 2005;**45**(3):531-39.

183. Huerta C, Zhao SZ, García Rodríguez LA. Risk of acute liver injury in patients with diabetes. *Pharmacotherapy* 2002;**22**(9):1091-96.

184. Johansson S, Wallander M-A, de Abajo FJ, García Rodríguez LA. Prospective drug safety monitoring using the UK primary-care General Practice Research Database: Theoretical framework, feasibility analysis and extrapolation to future scenarios. *Drug Safety: An International Journal Of Medical Toxicology And Drug Experience* 2010;**33**(3):223-32.

185. Khan NF, Perera R, Harper S, Rose PW. Adaptation and validation of the Charlson Index for Read/Oxmis coded databases. *BMC Fam Pract* 2010;**11**:1-1.

186. Langan SM, Groves RW, Card TR, Gulliford MC. Incidence, mortality, and disease associations of *Pyoderma gangrenosum* in the United Kingdom: A retrospective cohort study. *The Journal Of Investigative Dermatology* 2012;**132**(9):2166-70.

187. Lawrenson R, Todd JC, Leydon GM, Williams TJ, Farmer RD. Validation of the diagnosis of venous thromboembolism in general practice database studies. *Br J Clin Pharmacol* 2000;**49**(6):591-96.

188. Lawson DH, Sherman V, Hollowell J. The General Practice Research Database. Scientific and ethical advisory group. *QJM: Monthly Journal Of The Association Of Physicians* 1998;**91**(6):445-52.

189. Padwal RS, Klarenbach SW, Wang X, Sharma AM, Karmali S, Birch DW, et al. A simple prediction rule for all-cause mortality in a cohort eligible for bariatric surgery. *JAMA Surgery* 2013;**148**(12):1109-15.

190. Smeeth L, Hall AJ, Fombonne E, Rodrigues LC, Huang X, Smith PG. A case-control study of autism and mumps-measles-rubella vaccination using the General Practice Research Database: Design and methodology. *BMC Public Health* 2001;**1**:2-2.

191. Soriano JB, Maier WC, Visick G, Pride NB. Validation of general practitioner-diagnosed COPD in the UK General Practice Research Database. *Eur J Epidemiol* 2001;**17**(12):1075-80.

192. Tannen RL, Weiner MG, Xie D. Replicated studies of two randomized trials of angiotensin-converting enzyme inhibitors: Further empiric validation of the 'prior event rate ratio' to adjust for unmeasured confounding by indication. *Pharmacoepidemiol Drug Saf* 2008;**17**(7):671-85.

193. Thomas SL, Edwards CJ, Smeeth L, Cooper C, Hall AJ. How accurate are diagnoses for rheumatoid arthritis and juvenile idiopathic arthritis in the General Practice Research Database? *Arthritis Rheum* 2008;**59**(9):1314-21.

194. Van Staa TP, Abenhaim L, Cooper C, Zhang B, Leufkens HG. The use of a large pharmacoepidemiological database to study exposure to oral corticosteroids and risk of fractures: Validation of study population and results. *Pharmacoepidemiol Drug Saf* 2000;**9**(5):359-66.

195. Kim JB, Sells P, Publications C. *English syntax: An introduction*: CSLI Publications, 2008.

196. Strumpf M, Douglas A. *The grammar bible*. New York: St. Martin's Press, 2004.

197. Office for National Statistics. Cancer registration statistics, England, 2011. Secondary Cancer registration statistics, England, 2011. 2013. http://www.ons.gov.uk/ons/dcp171778_315795.pdf.