

## Previous versions

Version 1: BMJ Open-2010-000025

# Using free text information to explore misclassification in dating diagnoses of ovarian cancer. Observational study using the General Practice Research database.

A. Rosemary Tate<sup>1</sup>, senior research fellow,  
Alexander G. R. Martin<sup>2</sup>, specialist registrar, oncology,  
Aishath Ali<sup>1</sup>, research assistant, Jackie A. Cassell<sup>1</sup>, professor.  
<sup>1</sup> Brighton and Sussex Medical School, Falmer, Brighton BN1 9PH, UK.  
<sup>2</sup> Barts and the London NHS Trust, London, UK.

Correspondence to A. R. Tate, [rosemary.tate@bsms.ac.uk](mailto:rosemary.tate@bsms.ac.uk), 44 (0)1273 644562

## Abstract

**Objectives:** Primary care databases provide a unique resource for health care research, but most researchers currently use only the Read codes indicating diagnosis, symptoms etc., ignoring (much harder to access) information in the free text. We investigate how much information on ovarian cancer diagnosis is "hidden" in the free text and how this affects the coded date of diagnosis.

## Design:

Anonymised free text records from the General Practice Research Database of 344 women with a Read code indicating ovarian cancer between 01/06/2002 and 31/05/2008 were used to compare the date the diagnosis was first coded with the date the diagnosis was recorded in the free text. Text relating to a diagnosis were identified a) from the date of coded diagnosis and b) by searching for words relating to the ovary.

## Results :

90% of cases had information relating to their ovary in free text; 45% indicating a definite diagnosis of ovarian cancer. 22% had text confirming a diagnosis prior to the coded date; 10% over 4 weeks before. Four patients did not have ovarian cancer and 10% had only ambiguous or suspected diagnoses associated with the ovarian cancer code.

## Conclusions

There was a vast amount of extra information relating to diagnoses in the free text. Although in most cases text confirmed the diagnosis, it also showed that in some cases GPs do not code a definite diagnosis on the date that it is confirmed. For diseases which rely on hospital consultants for diagnosis, free text (particularly letters) is invaluable, not only for accurate dating of diagnosis and referrals, but also for identifying misclassified cases.

## Article Summary

### Article focus

- How much information on ovarian cancer diagnosis is "hidden" in the free text of primary care records?
- How accurate is the date of diagnosis based on Read codes?
- How many cases might be misclassified if codes alone are used to identify diagnoses?

### Key messages

- Free text contains much extra information on ovarian cancer diagnoses.
- This information can be used to more accurately date diagnosis and to identify cases that have not been coded.

- For certain disease areas, particularly where a specialist care is involved, free text should be used to determine the extent of misclassification associated with both the date of diagnosis and identification of cases.

#### Strengths and limitations of this study

- An in-depth analysis of information relating to ovarian cancer diagnoses using free text records from a large primary care database.
- We did not have access to letters that had been scanned in as images, so will have missed some important information.
- We ignored text that did not explicitly refer to the patient's ovaries and thus did not investigate pathways of care or symptoms. This is the topic of a separate study which is already underway

#### Introduction

UK primary care databases provide a valuable source of information for research on disease epidemiology, drug safety and adverse drug reactions. The records in these databases contain a mix of coded and free text data. Diseases, symptoms and clinical events are coded using 'Read' codes enabling searching for clinical entities. Analyses of existing large-scale electronic patient records - most extensively collated in the form of large primary care datasets such as the General Practice Research Database (GPRD), The Health Improvement Network (THIN) and QResearch - have almost exclusively exploited coded data. Such data are readily accessible to the classical methods of epidemiological analysis, once the complexities of defining and selecting a patient cohort have been overcome.

However, since clinicians can choose whether and how to code a consultation, an unknown amount of clinical data is in effect "hidden" outside the coded data in the free text notes. Free text records may contain further information on diagnosis (e.g. in the hospital letters) and are also likely to contain important information on the severity of symptoms or on additional symptoms which have not been coded. The degree to which clinical information is coded, and how this varies between by practitioner, practice, or type of clinical problem is currently unknown. The aim of this study is to determine how much extra information on ovarian cancer diagnosis is recorded in the free text, how often this is recorded before the coded date of diagnosis, and whether information from free text is needed for more accurate dating of diagnosis in research studies.

#### Methods

##### Data

This study builds on previous work on dating of diagnosis [1] where we used coded records from the General Practice Research database; one of the largest primary care databases in the UK. The GPRD contains anonymised longitudinal data on a representative sample of the UK population.

Records are being collected on over 4 million active patients who are registered for care in general practice from around 500 primary care practices throughout the UK. These records are created during consultations or processing correspondence, and are widely used in research on disease epidemiology, drug safety and adverse drug reactions. [2] The target population consisted of all females between 40 and 80 years of age (inclusive) who were alive and registered with a GPRD contributing practice on June 1, 2002. From this population, all women with an incident diagnosis of ovarian cancer recorded during June 1, 2002 - May 31, 2007 (i.e with a Read codes: B440.00 (Malignant neoplasm of ovary) B440.11 (Cancer of ovary ) or B44..00 (Malignant neoplasm of ovary and other uterine adnexa) were identified (n=1107). From these we selected 374 patients by randomly selecting one third of the contributing practices. Of these, 344 patients were used for this study after excluding 3 cases with a prior ambiguous diagnosis before the study period and 27 patients who had been registered with the GP for less than 2 years before diagnosis. Full details of the sample selection procedures are provided in [1].

At each consultation the GP is required to enter one or more Read codes into the computer system and, for each code entered, is given the option to add text which will be associated with that code. Read codes were developed in the 1980s and are currently used for coding clinical events in primary care in the United Kingdom. Each code has an associated text description e.g "Abdominal Pain", "Ovarian Cancer", "Letter from specialist", which are available on GP systems (usually as a drop down

menu) to help them record the correct code. The GP also has the option to enter a date indicating when the event occurred the "event date", if this differs from the date of the consultation. In this paper unless otherwise stated "date" will refer to the event date.

For this present study we obtained anonymised free text records for all the consultations recorded during the period 12 months prior to the date of the earliest Read code indicating a referral for or suspicion of ovarian cancer (Date 4) and up to and including the date of definite diagnosis (Date1) where the dating scheme follow the definitions of our earlier paper i.e.

**Date 1. Date of definite diagnosis** Earliest recorded date of definite diagnostic code.

**Date 2. Date of ambiguous diagnosis** Date 1, or, if present, the first date of an "ambiguous" code (e.g. "Cancer" "Carcinomatosis" with no previous cancer diagnosed in another site) if this occurs prior to but within two years of Date 1.

**Date 3. Date GP first knew, or suspected a diagnosis** Date 2, or, if present, first date of code indicating GP already knew of a cancer diagnosis if this occurs prior to but within two years of Date 2. (e.g cancer care review "chemotherapy" with no previous cancer diagnosis )

**Date 4. Date of first suspicion of, or first referral for ovarian cancer** Date 3, or, if present, the First date of a code for an investigation or referral to a gynaecologist if this is earlier than but within 12 months of Date 3. Dates 1 and 4 were different in 73% of the 344 ovarian cancer cases (67% of cases had tests or referrals prior to the diagnosis date). Full details of the codes which were used to define these dates are given in our earlier paper. [1]

#### **Extraction of information on diagnosis from the free text records**

In order to find information on ovarian cancer diagnoses in the free text, all free text records that referred to the ovary were identified by automatically extracting records containing the fragment "ovar", "ov " , or "ov." (in either upper or lower case). A manual inspection of the results revealed that all the matching strings referred to the ovary except for two referring to the drug "Novartin". These records were excluded as were 6 records that referred to a family history of ovarian cancer, i.e. in the patient's mother or sister. Texts recorded on the date of coded diagnosis (i.e. text that was associated with the Read code for ovarian cancer or other Read codes recorded on the same date) were then merged with these records. Text records, together with their associated Read codes, were grouped chronologically by patient id and where there were a number of text records for a patient on the same (event) date these texts were counted as a single record.

#### **Classification scheme for the free text**

The first 50 text records and their associated Read code descriptions were then examined by a gynaecological oncologist (AM). Since the major purpose of this study was to determine how often the GP recorded a definite diagnosis prior to coding it, we decided to err on the side of caution when classifying cases as "definite".

A scheme for classifying a diagnosis recorded in the text records was developed as follows:

1. *Blank*: there is no information in the text relating to diagnosis of either a benign or malignant condition, or that ovarian cancer is suspected.
2. *Benign*: text indicating definite diagnosis of a benign condition eg ovarian (cyst)adenoma.
3. *Suspected*: text indicating that ovarian cancer is suspected but with no definite diagnosis yet, e.g. "possible" or "probable" "highly likely". Alternatively, a surgeon's report (or GP entry relating to report) may describe appearance of ovaries, presence of peritoneal spread, ascites etc. Although surgeons can sometimes be very confident of the diagnosis based on operative appearances, and blood tests/radiology investigations and state the diagnosis as "ovarian cancer", we classify text as a suspected cancer if there was no mention of histological or cytological confirmation.

4. *Ambiguous*: an ambiguous diagnosis (e.g. "tumour", "metastatic cancer") which could be primary or secondary or a cancer other than ovarian.

5. *Secondary*: where the subject of the text (and Read code) is a documented primary malignancy of non-ovarian origin.

6. *Definite*: text indicating that a diagnosis of ovarian cancer has been confirmed. This confirmation would have been made by a histological or cytological confirmation of ovarian cancer, for example after surgery such as laparotomy, or cytology from ascitic fluid drainage. In the cases where the information relating to how the doctor came up with the diagnosis is not there e.g. simply "ovarian cancer" we assume that it has been confirmed appropriately. We included borderline ovarian cancers in this category, although they are classed as semi-malignant.

7. *Negated*: text indicating specifically that there is no ovarian cancer (in spite of the Read Code)  
The full set of selected text records were then inspected and assigned a provisional classification by the first author (ART). These were checked by AM who reassigned any that had been incorrectly classified or which were not clear cut enough for a non-specialist to classify. Text that had been recorded on the date of diagnosis was double-checked to see whether or not it confirmed the Read code for a diagnosis. Any Read codes for ovarian cancer which had no associated text, or text not relevant to the diagnosis, were assumed to be correct. In addition, each text record was classified as either a "letter" or "GP notes" and if there was information on the stage or grade this was recorded.

## Results

The total number of text records in the specified time period was 7860 representing 5777 consultations for 340 of the 344 patients. The median number of text records per patient was 19. Of these, 678 text records (representing 245 patients) were found to contain a reference to the patient's ovary. When these were merged with text recorded on the same date as the coded diagnosis the number of text records increased to 1007, representing 311 patients. The total number of text records, after combining text recorded on the same event date for each patient (and discarding any "blank" text recorded on the date of diagnosis) was 706 (for 282 patients), 462 of which were recorded prior to the date of coded diagnosis (191 patients). The analysis was based on these 706 records.

Table 1: Typical scenarios and use of text to describe the diagnostic process for patients with ovarian cancer and number of texts and patients with text (referring to the ovary) according to our classification scheme

Scenario	Example	Classification	Texts	Patients
Suggestive clinical feature but no mention of ovarian malignancy in text (prior to coded date)	Seen by GP. Suspected ovarian cyst or ovarian mass e.g "lump ? ovary" but no mention of cancer	Blank	151*	105 (31%)
Specific statement that clinical feature is benign.	Text states that cyst or lump is benign or that there is no evidence (yet) of malignancy e.g. "multiple fibroids", "thought to have ovarian ca but histology so far has shown benign cyst", "The curettings were benign"	Benign	7	7 (2%)
Referred for investigation of possible ovarian cancer	Seen by GP. Symptoms and signs suspicious of possible ovarian cancer, so referred for urgent	Suspected	116	85 (25%)

	so referred for urgent scan/blood tests or gynaecology outpatients appointment.			
Diagnostic test indicates suspicion of ovarian cancer	Has had scan/blood tests and report (or GP entry relating to report) is suspicious of ovarian cancer.			
Specialist's communication states that ovarian cancer very likely.	Has been seen at gynaecology outpatient clinic, and consultant letter (or GP entry relating to letter) may state ovarian cancer diagnosis very likely.			
Specialist communication following surgery indicates presumptive ovarian cancer, but histology/cytology awaited.	Has had surgery for probable ovarian cancer. Surgeon's report (or GP entry relating to report) may describe, <b>for example appearance of ovaries, presence of peritoneal spread, ascites.</b> The surgeon may be confident of a diagnosis, but is still awaiting a histological (or cytological) confirmation.			
A malignancy is suspected or confirmed, but site/origin not yet established.	Cancer is suspected, or has been confirmed but the site of the cancer has not yet been established.	Ambiguous	50	36 (10%)
Metastatic cancer (non-ovarian origin)	Cancer is from another site e.g. "metastatic lobular carcinoma of the breast".	Secondary	5	1 (0%)
Histologically or cytologically confirmed ovarian cancer	Histological or cytological confirmation of ovarian cancer has been made, e.g. from surgery such as laparotomy, or cytology from ascitic fluid drainage.	Definite	374	220 (64%)
Text provides additional confirmatory evidence of ovarian cancer (e.g.	The Read code for ovarian cancer is supplemented by extra			

grade, spread)	information in the text such as "grade III" , "sig metastatic spread", "chemotherapy"			
No further information available in free text on basis of the coded ovarian cancer diagnosis	Sometimes the information relating to how the doctor has come up with the diagnosis is not there e.g. simply "ovarian cancer".			
Text on date of diagnosis excludes ovarian cancer as a diagnosis (e.g. emendations indicating error, or discussion about cancer rather than diagnosis)	e.g. "this consultation has been changed as wrong diagnosis entered" , "Pt very concerned about possible cancer of ovary.. ... healthy eating and exercise discussed."	Negation	3	3 (1%)

\* "Blank" text that was recorded on the date of diagnosis was not included in this category

After examining the text records it was clear that information regarding possible ovarian cancer can be recorded at several different stages of the diagnostic process (Table 1). Approximately 25% of the texts appeared to be electronic letters. These were much more detailed than the GP notes which were often quite terse with misspellings and abbreviations, as illustrated by some of the examples in Table 1. However, not all the letters were available as electronic text; many of the records with a Read code indicating a letter just reported the result of the letter, and approximately 5% of texts indicated that a letter was available elsewhere (e.g as a scanned letter on an image viewer).

According to our classification scheme, 64% of patients had text either recording or confirming a definite diagnosis and 32% had a "probable" or "ambiguous" diagnosis. Figure 1 depicts the text classifications in relation to the coded date of diagnosis.

#### Text classified as "Definite"

The majority of text records that were classified as definite were recorded on the same date as their first code (205 (60%) patients). However 74 (22%) patients had a "definite" diagnosis recorded in the text prior to this date. The median (IQR) difference between the date that the diagnosis was coded and the date of the text for these 74 patients was 24 [8,67] days, with 34 (10%) of patients having a diagnosis more than 4 weeks before. Six patients had text stating that they had a recurrence of previously diagnosed ovarian cancer. All 6 of these had a prior definite diagnosis in the free text and one had a previous ambiguous code "carcinoma in situ of the ovary", a year earlier.

Fourteen patients were classified as definite prior to the derived date of first suspicion or investigation for ovarian cancer (date 4), 10 when this differed from date 1, but only 4 of these (of which 3 were described in the text as a recurrent case) had a diagnosis in the text more than 4 weeks prior to date 4.

Approximately two thirds of the texts indicating a definite diagnosis had been entered in association with a Read code that was not ovarian cancer (including 55 texts recorded on date 1) . Approximately one third of these had a code indicating a visit to an oncologist or gynaecologist or a "letter from specialist" or similar. The other third were associated with a variety of different codes, for example a code for an operation or hospital discharge letter, a generic cancer code (e.g "Adenocarcinomas") or

a code which bore no relation to the diagnosis at all - e.g. "Excepted from diabetes quality indicators" (4 patients), "Fracture of neck of femur" (1 patient).

Twenty-eight percent of patients had information on the stage, grade or spread of the tumour.

### Other classifications

39 patients (11%) were classified as not having a definite diagnosis of ovarian cancer on or prior to the date of coded diagnosis (Figure 1). The text records of 4 patients indicated that there was definitely no ovarian cancer: one had a metastasis in the breast, two a cancer in another site and one was worried she might have it, but subsequent coded records for this patient showed no further indication of cancer, so we assumed this had been coded in error. The other 35 patients had only a suspected or ambiguous diagnosis. Examination of subsequent codes for these patients indicated that all these patients did indeed have cancer (19 died or went into a hospice, 15 had a subsequent (repeated) code for ovarian cancer and most had codes for chemotherapy or cancer care).

31 patients had an ambiguous, or suspected diagnosis before date 4, (17 where date 1 and date 4 were different); the median difference was 13 [6,25] days, with 5 (3) patients having this classification more than 4 weeks prior.

### Discussion

This analysis of GPRD records of ovarian cancer patients shows that there is a large amount of free text differed associated with the GP records which contains a great deal of extra information on the patient's diagnosis, including information on when the diagnosis was notified to the GP and the stage, grade and spread of the tumour (for 28% of patients). In approximately 25% of cases this information was contained in an electronic copy of a letter from a specialist or hospital discharge. 282 (82%) of the 344 patients had free text relating to their ovarian pathology with 220 of these having text which recorded or confirmed a diagnosis of ovarian cancer.

74/344 (22%) of all patients had definite indication of ovarian cancer in the free text prior to the first code, with half more than 24 days before. However, there were only 10 definite diagnoses in the text before our derived date for first suspicion of or referral for ovarian cancer, which we argue [1] is a better date for investigating primary care delays. A number of other inconsistencies were corrected by reference to the free text. Four patients did not have ovarian cancer at all and 6 were recurrences of a much earlier ovarian cancer not evident from the codes. Furthermore 35 were classified as having only a suspected or ambiguous diagnosis (but not definite) on the date of the code.

There is no Read code for a possible, probably or highly likely diagnosis, and this may explain why some of the patients had only a suspected or ambiguous diagnosis associated with a definite code. The delay between the GP recording the diagnosis in the text and coding might be partly explained by the administrative practices of the surgery where, for example, practice staff code and date the letters when they arrive and the GPs fills in the detail at a later date. This supposition was supported by the fact that only about one quarter of the "definite" diagnoses recorded on the date of diagnosis came from letters whereas nearly half of "definite" texts recorded before that date were letters.

The text records show that even when definite diagnosis codes were assigned, the decision was not always clear cut. It also showed that diagnosis in the text were often associated with a generic Read code. Since most studies of diseases are based on the codes, uncoded cases will lead to incorrect estimates of the incidence of the disease and will also have an impact on case selection.

To our knowledge this is only work to explicitly explore and report dating of diagnosis in GP records using free text. Other studies have used creative use of code lists for example our previous study [1] or GP questionnaires e.g. [3] to investigate the accuracy of the coded date. In our previous study [1], which was based on the same data as this study, we found that 12% of cases had an ambiguous code prior to the coded date and that 19% had a code indicating a possible prior diagnosis. There have been a handful of studies using free text to verify clinical conditions in combination with codes, a few of them have used free text to identify cases [4, 5] while the majority used free text to verify and validate coded information.[6-8, 3, 9, 6]

However, most of these works have little or no detail on the process of selection and review of the free text used within the studies. A notable exception is a validation study on congenital heart defects by Wurtst et al. [7] which provides the free text selection criteria, together with the proportion of the

data containing free text information, the level of detail of the free text and how much additional information (in this case, the type and size of a ventricular septal defect) is available in the free text.

In this study we only looked at cases that had a definite diagnosis code for ovarian cancer, so we will have missed any cases that only had a diagnosis in the free text (false negatives). Our recent comparison with the cancer registries [10] indicated that around 9% of cases may not have been coded and thus missed this way. Another limitation was that we did not have access to letters that were not available in electronic text format which may have contained important information that was not relayed by the GP. We estimate (data not shown) that approximately 20-25% of coded hospital letters for these patients will not have any information on their content (either a code or text) entered on the date that they were received.

### **Conclusions**

This study gives an in-depth insight on the extra information that is contained in the free text part of the records relating to a suspected or confirmed diagnosis of ovarian cancer. A large amount of information in free text is available that modifies the date or reveals incorrect classification as a case even for "hard" outcomes such as ovarian cancer that is considered well documented in primary care records. This shows that a) the quality of information in primary care records is better than one might think, but b) free text needs to be routinely explored to take advantage of this quality information. It is likely that the proportion of information concealed in free text will be higher for less "hard" outcomes in certain disease areas. We are therefore working with Natural Language Processing experts to find ways of extracting relevant information from large volumes of text.

### **Acknowledgements**

We would like to thank the PREP team and John Parkinson for helpful comments and discussions.

### **Author's Contributions**

ART conceived and wrote the paper, read and classified the free text and carried out the subsequent analysis. AM devised the classification scheme, read through and classified the free text, wrote part of the paper and provided expert advice. JC was involved in the conception of and writing the paper. AA participated in writing the paper (including the literature review) and assisted with data management. All authors had full access to all of the data (including statistical reports and tables) in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis. ART is the guarantor.

### **Data sharing**

Codelists and statistical code available from the corresponding author.

### **Funding**

This work was supported by the Wellcome Trust [086105/Z/08/Z]. Access to the GPRD database was funded through the Medical Research Councils licence agreement with MHRA. The authors were independent from the funder and sponsor, who had no role in conduct, analysis or the decision to publish. This study is based in part on data from the Full Feature General Practice Research Database obtained under licence from the UK Medicines and Healthcare Products Regulatory Agency. However, the interpretation and conclusions contained in this study are those of the author/s alone.

### **1 Ethics Approval**

Access to the dataset was approved by the Independent Scientific Advisory Committee (Protocol 07 069).

### **Competing Interests**

All authors declare that the answer to the questions on your competing interest form are all No and therefore have nothing to declare.

### **References**

[1] Tate, A. R., Martin, A. G. R., Murray-Thomas, T., Anderson, S. R., and Cassell, J. A. (2009). 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 Medical Research Methodology* 9.

- [2] GPRD (2010). GPRD. Excellence in public health research. <http://www.gprd.com>.
- [3] Hammad, T. A., McAdams, M. A., Feight, A., Iyasu, S., and Dal Pan, G. J. (2008). Determining the predictive value of Read/OXMIS codes to identify incident acute myocardial infarction in the General Practice Research Database. *PHARMACOEPIDEMIOLOGY AND DRUG SAFETY* 17, 1197-1201.
- [4] Masso Gonzalez, E. L., Johansson, S., Wallander, M.-A., and Garcia Rodriguez, L. A. (2009). Trends in the prevalence and incidence of diabetes in the UK: 1996-2005. *JOURNAL OF EPIDEMIOLOGY AND COMMUNITY HEALTH* 63, 332-336.
- [5] Martinez, C., Assimes, T. L., Mines, D., Dell'Aniello, S., and Suissa, S. (2010). Use of venlafaxine compared with other antidepressants and the risk of sudden cardiac death or near death: a nested case-control study. *BRITISH MEDICAL JOURNAL* 340.
- [6] Stowe, J., Andrews, N., Wise, L., and Miller, E. (2009). Investigation of the Temporal Association of Guillain-Barre Syndrome With Influenza Vaccine and Influenza-like Illness Using the United Kingdom General Practice Research Database. *AMERICAN JOURNAL OF EPIDEMIOLOGY* 169, 382-388.
- [7] Wurst, K. E., Ephross, S. A., Loehr, J., Clark, D. W., and Guess, H. A. (2007). The utility of the general practice research database to examine selected congenital heart defects: a validation study. *Pharmacoepidemiology Drug Safety* 16, 867-877.
- [8] Ruigomez, A., Martin-Merino, E., and Garcia Rodriguez, L. A. (2010). Validation of ischemic cerebrovascular diagnoses in the health improvement network (THIN). *PHARMACOEPIDEMIOLOGY AND DRUG SAFETY* 19, 579-585.
- [9] Charlton, R. A., Weil, J. G., Cunnington, M. C., and de Vries, C. S. (2010). 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 SAFETY* 33, 741-750.
- [10] Tate, A. R., Nicholson, A. C., and Cassell, J. A. (2010). Are GPs under-investigating older patients presenting with symptoms of ovarian cancer? Observational study using General Practice Research Database. *British Journal of Cancer* 102, 947-951.

# Using free text information to explore how and when GPs code a diagnosis of ovarian cancer. Observational study using the General Practice Research database.

A. Rosemary Tate<sup>1</sup>, senior research fellow,  
Alexander G. R. Martin<sup>2</sup>, specialist registrar, oncology,  
Aishath Ali<sup>1</sup>, research assistant, Jackie A. Cassell<sup>1</sup>, professor.  
<sup>1</sup> Brighton and Sussex Medical School, Falmer, Brighton BN1 9PH, UK.  
<sup>2</sup> Barts and the London NHS Trust, London, UK.

Correspondence to A. R. Tate, r.tate@bsms.ac.uk, 44 (0)1273 644562

## Abstract

**Objectives:** Primary care databases provide a unique resource for health care research, but most researchers currently use only the Read codes for their studies, ignoring (much harder to access) information in the free text. We investigate how much information on ovarian cancer diagnosis is "hidden" in the free text **and the time lag between a diagnosis being described in the text or in a hospital letter and the patient being given a READ code for that diagnosis.**

**Design:** Anonymised free text records from the General Practice Research Database of 344 women with a Read code indicating ovarian cancer between 01/06/2002 and 31/05/2008 were used to compare the date the diagnosis was first coded with the date the diagnosis was recorded in the free text. **Free text** relating to a diagnosis was identified a) from the date of coded diagnosis and b) by searching for words relating to the ovary.

**Results:** 90% of cases had information relating to their ovary in free text. 45% had text indicating a definite diagnosis of ovarian cancer. 22% had text confirming a diagnosis prior to the coded date; 10% over 4 weeks before. Four patients did not have ovarian cancer and 10% had only ambiguous or suspected diagnoses associated with the ovarian cancer code.

**Conclusions** There was a vast amount of extra information relating to diagnoses in the free text. Although in most cases text confirmed the coded diagnosis, it also showed that in some cases GPs do not code a definite diagnosis on the date that it is confirmed. For diseases which rely on hospital consultants for diagnosis, free text (particularly letters) is invaluable, not only for accurate dating of diagnosis and referrals, but also for identifying misclassified cases.

## Article Summary

### Article focus

- How much information on ovarian cancer diagnosis is "hidden" in the free text of primary care records?
- How accurate is the date of diagnosis based only on Read codes?
- How many cases might be misclassified if codes alone are used to identify diagnoses?

### Key messages

- Free text contains much extra information on ovarian cancer diagnoses, including grade and stage of the cancer and the dates that the patient was investigated and diagnosed in secondary care.
- This information can be used to determine the date that a diagnosis was notified to the GP and to identify cases that have not been coded.

- For certain disease areas, particularly where a specialist care is involved, free text should be used to determine the extent of misclassification associated with both the (coded) date of diagnosis and identification of cases.

#### Strengths and limitations of this study

- An in-depth analysis of information relating to ovarian cancer diagnoses using free text records from a large primary care database.
- We did not have access to letters that had been scanned in as images, so will have missed some important information.
- We only looked at cases which had been assigned an unambiguous Read code for ovarian cancer and thus will have missed cases with no code or an ambiguous code.
- We ignored text that did not explicitly refer to the patient's ovaries and thus did not investigate pathways of care or symptoms. This is the topic of a separate study which is already underway.
- We have only looked at ovarian cancer, and cannot say whether our findings will generalise to other diseases.

#### Introduction

UK primary care databases provide a valuable source of information for research on disease epidemiology, drug safety and adverse drug reactions. The records in these databases contain a mix of coded and free text data. Diseases, symptoms and clinical events are coded using 'Read' codes enabling searching for clinical entities. Analyses of existing large-scale electronic patient records - most extensively collated in the form of large primary care datasets such as the General Practice Research Database (GPRD), The Health Improvement Network (THIN) and QResearch - have almost exclusively exploited coded data. Such data are readily accessible to the classical methods of epidemiological analysis, once the complexities of defining and selecting a patient cohort have been overcome.

However, since clinicians can choose whether and how to code a consultation, an unknown amount of clinical data is in effect "hidden" outside the coded data in the free text notes. Free text records, **as distinct from coded records**, may contain further information on diagnosis (which have been copied or imported from hospital letters) and are also likely to contain important information on the severity of symptoms or on additional symptoms which have not been coded. The degree to which clinical information is coded, and how this varies between by practitioner, practice, or type of clinical problem is currently unknown.

The aim of this study is to determine how much extra information on ovarian cancer diagnosis is recorded in the free text, how often this is recorded before the event date for which the diagnosis is coded, and whether information from free text is needed for more accurate dating of diagnosis in research studies.

#### Methods

##### Data

This study builds on previous work on dating of diagnosis [1] where we used coded records from the General Practice Research database; one of the largest primary care databases in the UK. The GPRD contains anonymised longitudinal data on a representative sample of the UK population. Records are being collected on over 4 million active patients who are registered for care in general practice from around 500 primary care practices throughout the UK. These records are created during consultations or processing correspondence, and are widely used in research on disease epidemiology, drug safety and adverse drug reactions. [2]

The target population consisted of all females between 40 and 80 years of age (inclusive) who were alive and registered with a GPRD contributing practice on June 1, 2002. From this population, all women with an incident diagnosis of ovarian cancer recorded during June 1, 2002 - May 31, 2007 (i.e with one of the Read codes: B440.00 (Malignant neoplasm of ovary) B440.11 (Cancer of ovary) or B44..00 (Malignant neoplasm of ovary and other uterine adnexa) were identified (n=1107). From these we selected 374 patients by randomly selecting one third of the contributing practices. Of these, 344 patients were used for this study after excluding 3 cases with a prior ambiguous diagnosis before

the study period and 27 patients who had been registered with the GP for less than 2 years before diagnosis. Full details of the sample selection procedures are provided in [1].

At each consultation the GP may enter one or more Read codes into the computer system and, for each code entered, is given the option to add free text which will be associated with that code. Read codes were developed in the 1980s and are currently used for coding clinical events in primary care in the United Kingdom. Each code has an associated text description e.g "Abdominal Pain", "Ovarian Cancer", "Letter from specialist", which are available on GP systems (usually as a drop down menu) to help them record the correct code. The GP also has the option to enter a date indicating when the event occurred the "event date", if this differs from the date of the consultation. In this paper unless otherwise stated "date" will refer to the event date.

For this present study we obtained anonymised free text records for all the consultations recorded during the period 12 months prior to the date of the earliest Read code indicating a referral for or suspicion of ovarian cancer (Date 4) and up to and including the date of definite diagnosis (Date1) where the dating scheme follow the definitions of our earlier paper i.e.

**Date 1. Date of first definite diagnosis - (referred to in this paper as date of coded diagnosis)** Earliest recorded (event) date of definite diagnostic code (**Read codes as above**).

**Date 2. Date of first ambiguous diagnosis** Date 1, or, if present, the first date of an "ambiguous" code (e.g. "Cancer" "Carcinomatosis" with no previous cancer diagnosed in another site) if this occurs prior to but within two years of Date 1.

**Date 3. Date GP first knew, or suspected a diagnosis** Date 2, or, if present, first date of code indicating GP already knew of a cancer diagnosis if this occurs prior to but within two years of Date 2. (e.g cancer care review "chemotherapy" with no previous cancer diagnosis)

**Date 4. Date of first suspicion of, or first referral for ovarian cancer** Date 3, or, if present, the First date of a code for an investigation or referral to a gynaecologist if this is earlier than but within 12 months of Date 3.

Dates 1 and 4 were different in 73% of the 344 ovarian cancer cases (67% of cases had tests or referrals prior to the diagnosis date). Full details of the codes which were used to define these dates are given in our earlier paper. [1]

### **Extraction of information on diagnosis from the free text records**

In order to find information on ovarian cancer diagnoses in the free text, all free text records that referred to the ovary were identified by automatically extracting records containing the fragment "ovar", "ov " , or "ov." (in either upper or lower case). A manual inspection of the results revealed that all the matching strings referred to the ovary except for two referring to the drug "Novartin". These records were excluded as were 6 records that referred to a family history of ovarian cancer, i.e. in the patient's mother or sister. **Textual data recorded on the date of coded diagnosis (i.e. text that was associated with the Read code for ovarian cancer or other Read codes recorded on the same date) were then merged with these records. Textual records, together with their associated Read codes, were grouped chronologically by patient id and where there were a number of free text records for a patient on the same (event) date these were counted as a single record.**

### **Classification scheme for the free text**

The first 50 text records and their associated Read code descriptions were then examined by a gynaecological oncologist (AM). Since the major purpose of this study was to determine how often the GP recorded a definite diagnosis prior to coding it, we decided to err on the side of caution when classifying cases as "definite". A scheme for classifying a diagnosis recorded in the text records was developed as follows:

1. *Blank*: there is no information in the text relating to diagnosis of either a benign or malignant condition, or that ovarian cancer is suspected.

2. *Benign*: text indicating definite diagnosis of a benign condition eg ovarian (cyst) adenoma.
3. *Suspected*: text indicating that ovarian cancer is suspected but with no definite diagnosis yet, e.g. "possible" or "probable" "highly likely". Alternatively, a surgeon's report (or GP entry relating to report) may describe for example, the appearance of ovaries, presence of peritoneal spread, or ascites, which implies suspicion of ovarian cancer. Although surgeons can sometimes be very confident of the diagnosis based on operative appearances, and blood tests/radiology investigations and state the diagnosis as "ovarian cancer", we classify text as a suspected cancer if there was no mention of histological or cytological confirmation.
4. *Ambiguous*: an ambiguous diagnosis, e.g. "tumour", which could be benign or could be a primary or secondary cancer, or "metastatic cancer" which could be a primary or secondary ovarian cancer or another type of cancer.
5. *Secondary*: where the subject of the text (and Read code) is a documented primary malignancy of non-ovarian origin.
6. *Definite*: text indicating that a diagnosis of ovarian cancer has been confirmed. This confirmation would have been made by a histological or cytological confirmation of ovarian cancer, for example after surgery such as laparotomy, or cytology from ascitic fluid drainage. In the cases where the information relating to how the doctor came up with the diagnosis is not there e.g. simply "ovarian cancer" we assume that it has been confirmed appropriately. We included borderline ovarian cancers in this category, although they are classed as semi-malignant.
7. *Negated*: text indicating specifically that there is no ovarian cancer (in spite of the Read Code)

The full set of selected text records were then inspected and assigned a provisional classification by the first author (ART). These were checked by AM who reassigned any that had been incorrectly classified or which were not clear cut enough for a non-specialist to classify. Text that had been recorded on the coded date of diagnosis was double-checked to see whether or not it confirmed the Read code for a diagnosis. Any Read codes for ovarian cancer which had no associated text, or text not relevant to the diagnosis, were assumed to be correct. In addition, each text record was classified as a either a "letter" or "GP notes" and if there was information on the stage or grade this was recorded.

## Results

The total number of text records, **in the specified time period, was 7860**, representing 5777 consultations for 340 of the 344 patients. The median number of text records per patient was 19. Of these, 678 text records (representing 245 patients) were found to contain a reference to the patient's ovary. When these were merged with text recorded on the same date as the coded diagnosis the number of text records increased to 1007, representing 311 patients. The total number of text records, after combining text recorded on the same event date for each patient (and discarding any "blank" text recorded on the date of diagnosis) was 706 (for 282 patients), 462 of which were recorded prior to the date of coded diagnosis (191 patients). The analysis was based on these 706 records.

After examining the text records it was clear that information regarding possible ovarian cancer can be recorded at several different stages of the diagnostic process (Table 1). Approximately 25% of the textual data appeared to be electronic letters. These were much more detailed than the GP notes which were often quite terse with misspellings and abbreviations, as illustrated by some of the examples in Table 1. However, not all the letters were available as electronic text; many of the records with a Read code indicating a letter just reported the result of the letter, and approximately 5% of free texts indicated that a letter was available elsewhere (e.g as a scanned letter on an image viewer).

According to our classification scheme, 64% of patients had text either recording or confirming a definite diagnosis and 32% had a "probable" or "ambiguous" diagnosis. Figure 1 depicts the text classifications in relation to the coded date of diagnosis.

Text classified as "Definite"

The majority of text records that were classified as definite were recorded on the same date as the patient's first ovarian cancer code (205 (60%) patients). However 74 (22%) patients had a "definite" diagnosis recorded in the text prior to this date. The median (IQR) difference between the date that the diagnosis was coded and the date of the text for these 74 patients was 24 [8,67] days, with 34 (10%) of patients having a diagnosis more than 4 weeks before. Six patients had text stating that they had a recurrence of previously diagnosed ovarian cancer. All 6 of these had a prior definite diagnosis in the free text and one had a previous ambiguous code "carcinoma in situ of the ovary", a year earlier.

Fourteen patients were classified as definite prior to the derived date of first suspicion or investigation for ovarian cancer (date 4), 10 when this differed from date 1, but only 4 of these (of which 3 were described in the text as a recurrent case) had a diagnosis in the text more than 4 weeks prior to date 4.

Approximately two thirds of the texts indicating a definite diagnosis had been entered in association with a Read code that was not ovarian cancer (including 55 texts recorded on date 1). Approximately one third of these had a code indicating a visit to an oncologist or gynaecologist or a "letter from specialist" or similar. The other third were associated with a variety of different codes, for example a code for an operation or hospital discharge letter, a generic cancer code (e.g "Adenocarcinomas") or a code which bore no relation to the diagnosis at all - e.g. "Excepted from diabetes quality indicators" (4 patients), "Fracture of neck of femur" (1 patient).

Twenty-eight percent of patients had information on the stage, grade or spread of the tumour.

### Other classifications

39 patients (11%) were classified as not having a definite diagnosis of ovarian cancer on or prior to the date of coded diagnosis (Figure 1). The text records of 4 patients indicated that there was definitely no ovarian cancer: one had a metastasis in the breast, two a cancer in another site and one was worried she might have it, but subsequent coded records for this patient showed no further indication of cancer, so we assumed this had been coded in error. The other 35 patients had only a suspected or ambiguous diagnosis. Examination of subsequent codes for these patients indicated that all these patients did indeed have cancer (19 died or went into a hospice, 15 had a subsequent (repeated) code for ovarian cancer and most had codes for chemotherapy or cancer care).

31 patients had an ambiguous, or suspected diagnosis before date 4, (17 where date 1 and date 4 were different); the median difference was 13 [6,25] days, with 5 (3) patients having this classification more than 4 weeks prior.

Table 1: Typical scenarios and use of text to describe the diagnostic process for patients with ovarian cancer and number of free texts and patients with text (referring to the ovary) according to our classification scheme

Scenario	Example	Classification	Texts	Patients
Suggestive clinical feature but no mention of ovarian malignancy in text (prior to coded date)	Seen by GP. Suspected ovarian cyst or ovarian mass e.g "lump ? ovary" but no mention of cancer	Blank	151*	105 (31%)
Specific statement that clinical feature is benign.	Text states that cyst or lump is benign or that there is no evidence (yet) of malignancy e.g. "multiple fibroids", "thought to have ovarian ca but histology so far	Benign	7	7 (2%)

	has shown benign cyst", "The curettings were benign"			
Referred for investigation of possible ovarian cancer	Seen by GP. Symptoms and signs suspicious of possible ovarian cancer, so referred for urgent scan/blood tests or gynaecology outpatients appointment.	Suspected	116	85 (25%)
Diagnostic test indicates suspicion of ovarian cancer	Has had scan/blood tests and report (or GP entry relating to report) is suspicious of ovarian cancer.			
Specialist's communication states that ovarian cancer very likely.	Has been seen at gynaecology outpatient clinic, and consultant letter (or GP entry relating to letter) may state ovarian cancer diagnosis very likely.			
Specialist communication following surgery indicates presumptive ovarian cancer, but histology/cytology awaited.	Has had surgery for probable ovarian cancer. Surgeon's report (or GP entry relating to report) may describe, <b>for example appearance of ovaries, presence of peritoneal spread, ascites.</b> The surgeon may be confident of a diagnosis, but is still awaiting a histological (or cytological) confirmation.			
A malignancy is suspected or confirmed, but site/origin not yet established.	Cancer is suspected, or has been confirmed but the site of the cancer has not yet been established.	Ambiguous	50	36 (10%)
Metastatic cancer (non-ovarian origin)	Cancer is from another site e.g. "metastatic lobular carcinoma of the breast".	Secondary	5	1 (0%)
Histologically or cytologically confirmed ovarian cancer	Histological or cytological confirmation of ovarian cancer has been made, e.g. from surgery such as	Definite	374	220 (64%)

	laparotomy, or cytology from ascitic fluid drainage.			
Text provides additional confirmatory evidence of ovarian cancer (e.g. grade, spread)	The Read code for ovarian cancer is supplemented by extra information in the text such as "grade III", "sig metastatic spread", "chemotherapy"			
No further information available in free text on basis of the coded ovarian cancer diagnosis	Sometimes the information relating to how the doctor has come up with the diagnosis is not there e.g. simply "ovarian cancer".			
Text on date of diagnosis excludes ovarian cancer as a diagnosis (e.g. emendations indicating error, or discussion about cancer rather than diagnosis)	e.g. "this consultation has been changed as wrong diagnosis entered", "Pt very concerned about possible cancer of ovary.. ... healthy eating and exercise discussed."	Negation	3	3 (1%)

\* "Blank" text that was recorded on the date of diagnosis was not included in this category

Figure 1: Number of patients and text classifications with event dates on and prior to the coded date of diagnosis

## Discussion

This analysis shows that primary care records hold a large amount of free text containing information on ovarian cancer diagnoses. The majority (82%) of the 344 patients had free text relating to their ovarian pathology and 64% had free text confirming an ovarian cancer code.

However, the information in the text was not always reflected in the codes. In some cases a "definite diagnosis" in the text field preceded the coded diagnosis, with 22% of patients having a "definite" classification of ovarian cancer recorded in the text prior to the date of the first ovarian cancer code. Half of these cases had a "definite" classification more than 24 days before the coded date, however, only 10 of these diagnoses occurred before our derived date for suspicion of, or referral for, ovarian cancer. A number of other inconsistencies were identified using the free text: 4 patients did not have ovarian cancer at all and 6 were recurrences of a much earlier ovarian cancer not evident from the codes.

The delay between the GP recording the diagnosis in the text and coding might be partly explained by incorrect entry of dates or by the administrative practices of the surgery where, for example, practice staff code and date the letters when they arrive and the GPs assigns the cancer code at a later date. This latter supposition was supported by the fact that only about one quarter of the "definite" classifications recorded on the date of diagnosis came from letters whereas nearly half of "definite" free texts recorded before that date were letters.

There is no Read code for a possible, probably or highly likely diagnosis, and this may explain why 10% of the patients classified as having only a suspected or ambiguous diagnosis nevertheless had an ovarian cancer code. Conversely, "definite" diagnoses in free text were often associated with a very general Read code (e.g. "letter from specialist"). Since most studies of diseases are based on the codes, wrongly or un-coded cases will lead to incorrect estimates of the incidence of the disease and will also have an impact on case selection. In addition, incorrectly entered dates of the notification of diagnosis will impact on the findings of studies on incidence of symptoms and delay prior to diagnosis. How much difference this will make will depend on the disease and also on the time period used for calculating incidence. For this dataset, redefining the (coded) date of diagnosis using the free text did not have much effect on estimates of delay or incidence of symptoms (data not shown), but this might not be the case for other diseases.

To our knowledge this is the only work to explicitly explore and report dating of diagnosis in GP records using the textual part of the record. Other studies have used creative use of code lists for example our previous study [1] or GP questionnaires e.g. [3] to investigate the accuracy of the coded date. There have been a handful of studies using free text to verify clinical conditions in combination with codes, a few of them have used free text to identify cases [4, 5] while the majority used free text to verify and validate coded information.[6-8, 3, 9, 6] However, with the exception of [7] most of these works have little or no detail on the process of selection and review of the free text that was used.

In this study we looked only at cases that had an unambiguous diagnosis code for ovarian cancer, so we will have missed any that had not been coded or which had ambiguous codes. Our recent comparison with the cancer registries [10] indicated that around 9% of cases may not have been coded and thus missed this way. Another limitation was that we did not have access to letters that were not available in electronic text format which may have contained important information that was not relayed by the GP. Letters from hospitals or specialists accounted for about 25% of free text records that we examined, but we did not have access to letters that were not available in electronic text format which may have contained important information that was not relayed by the GP. We estimate (data not shown) that approximately 20-25% of hospital letters for these patients will not have any information on their content (either a code or text) entered on the date that they were received. It is likely that many of these letters will contain information on diagnosis, particularly for cancer, where a specialist will always make the diagnosis. However, with the increased transfer of electronic records and even sharing of hospital records this problem is likely to be resolved in the near future.

## Conclusions

This study gives an in-depth insight on the extra information that is contained in the free text part of the records relating to a suspected or confirmed diagnosis of ovarian cancer. A large amount of information in free text is available that modifies the coded date or reveals incorrect classification as a case even for "hard" outcomes such as ovarian cancer that is considered well documented in primary care records. This shows that a) the quality of information in primary care records is better than one might think, but b) free text needs to be routinely explored to take advantage of this quality information. It is likely that the proportion of information concealed in free text will be higher for less "hard" outcomes in certain disease areas. We are therefore working with Natural Language Processing experts to find ways of extracting relevant information automatically (and therefore more cost effectively) from large volumes of text.

## Acknowledgements

We would like to thank the PREP team and John Parkinson for helpful comments and discussions.

## Author's Contributions

ART conceived and wrote the paper, read and classified the free text and carried out the subsequent analysis. AM devised the classification scheme, read through and classified the free text, wrote part of the paper and provided expert advice. JC was involved in the conception of and writing the paper. AA participated in writing the paper (including the literature review) and assisted with data management. All authors had full access to all of the data (including statistical reports and tables) in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis. ART is the guarantor.

## Data sharing

Codelists and statistical code available from the corresponding author.

## Funding

This work was supported by the Wellcome Trust [086105/Z/08/Z]. Access to the GPRD database was funded through the Medical Research Councils licence agreement with MHRA. The authors were independent from the funder and sponsor, who had no role in conduct, analysis or the decision to publish. This study is based in part on data from the Full Feature General Practice Research Database obtained under licence from the UK Medicines and Healthcare Products Regulatory Agency. However, the interpretation and conclusions contained in this study are those of the author/s alone.

## Ethics Approval

Access to the dataset was approved by the Independent Scientific Advisory Committee (Protocol 07 069).

## Competing Interests

All authors declare that the answer to the questions on your competing interest form are all No and therefore have nothing to declare

## References

- [1] Tate, A. R., Martin, A. G. R., Murray-Thomas, T., Anderson, S. R., and Cassell, J. A. (2009). 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 Medical Research Methodology* 9.
- [2] GPRD (2010). GPRD. Excellence in public health research. <http://www.gprd.com>.
- [3] Hammad, T. A., McAdams, M. A., Feight, A., Iyasu, S., and Dal Pan, G. J. (2008). Determining the predictive value of Read/OXMIS codes to identify incident acute myocardial infarction in the General Practice Research Database. *PHARMACOEPIDEMIOLOGY AND DRUG SAFETY* 17, 1197-1201.
- [4] Masso Gonzalez, E. L., Johansson, S., Wallander, M.-A., and Garcia Rodriguez, L. A. (2009). Trends in the prevalence and incidence of diabetes in the UK: 1996-2005. *JOURNAL OF EPIDEMIOLOGY AND COMMUNITY HEALTH* 63, 332-336.
- [5] Martinez, C., Assimes, T. L., Mines, D., Dell'Aniello, S., and Suissa, S. (2010). Use of venlafaxine compared with other antidepressants and the risk of sudden cardiac death or near death: a nested case-control study. *BRITISH MEDICAL JOURNAL* 340.
- [6] Stowe, J., Andrews, N., Wise, L., and Miller, E. (2009). Investigation of the Temporal Association of Guillain-Barre Syndrome With Inuenza Vaccine and Inuenzalike Illness Using the United Kingdom General Practice Research Database. *AMERICAN JOURNAL OF EPIDEMIOLOGY* 169, 382-388.
- [7] Wurst, K. E., Ephross, S. A., Loehr, J., Clark, D. W., and Guess, H. A. (2007). The utility of the general practice research database to examine selected congenital heart defects: a validation study. *Pharmacoepidemiology Drug Safety* 16, 867-877.
- [8] Ruigomez, A., Martin-Merino, E., and Garcia Rodriguez, L. A. (2010). Validation of ischemic cerebrovascular diagnoses in the health improvement network (THIN). *PHARMACOEPIDEMIOLOGY AND DRUG SAFETY* 19, 579-585.
- [9] Charlton, R. A., Weil, J. G., Cunnington, M. C., and de Vries, C. S. (2010). 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 SAFETY* 33, 741-750.

[10] Tate, A. R., Nicholson, A. C., and Cassell, J. A. (2010). Are GPs under-investigating older patients presenting with symptoms of ovarian cancer? Observational study using General Practice Research Database. *British Journal of Cancer* 102, 947-951.

Acceptance versions