SUPPLEMENTAL INFORMATION

Table of Contents

Patient Profiles

Profiles incorporated patients' entire history of Read codes (clinical diagnoses, surgical procedures, radiation therapy, and tumor morphology), relevant additional clinical information (e.g., prostate-specific antigen levels, mammograms, magnetic resonance images), and prescriptions, except those for the study overactive bladder drugs.

When Were Provisional Cases Considered Confirmed in the Patient Profile Review?

Provisional cases were confirmed when patient medical profiles presented supportive evidence of a cancer diagnosis, in particular, a relevant pathology (morphology) Read code, cancer-specific therapy (surgery, radiation therapy, chemotherapy, hormonal therapy) within 1 month before to 3 months after the cancer diagnosis code, or evidence of cancer care review by the general practitioner.[1, 2] Only surgical procedures that would be used to treat cancer were considered confirmatory (e.g., mastectomy was considered confirmatory, but a biopsy was not). Provisional cases were also confirmed if subsequent clinical events (referrals, hospitalizations, or death) were associated with clinical Read codes appropriate to the cancer diagnosis. Information contained in free-text comments was also taken into consideration for case confirmation.

Validation Process

Each patient's medical profile was reviewed by one physician from among the authors (JAK, JF, or AVM); results were entered into a "scorecard" (Excel spreadsheet) designed for the study. The scorecard included electronically populated fields for the patient identifier, cohort entry date, type of cancer (e.g., breast), cancer diagnosis date, and blank fields for the reviewer's decision regarding case status (with a pull-down menu for confirmed, provisional case, and non-case), reviewer-corrected type of cancer (if applicable), reviewer-corrected cancer diagnosis date (if applicable), and reviewer comments (if any). Reviewers corrected the date of diagnosis when there was clear indication that the cancer had been diagnosed before or after the code-based cancer diagnosis date and used clinical judgment, based on the information recorded in each provisional case's General Practitioner Online Database data, to determine the cancer type if there was any ambiguity.

To enhance consistency in evaluating provisional cases, all three reviewers initially examined and discussed 30 randomly selected profiles without free-text comments and an additional 20 with free-text comments. During the subsequent profile review process, provisional cases for which status was not clear were discussed, always including the specialist in medical oncology/hematology (JAK), until reaching consensus. After all reviews were completed, data recorded in the scorecards were imported into SAS and incorporated into the analytical file.

Request for Free-Text Comments

Criteria to Select Cancer Cases When Free-Text Comments Were Not Expected to Provide Information Beyond the Information Identified by Codes

- If the first occurrence of the Read code for "cancer care review" (8BAV.00) appears on any subsequent record after the initial cancer date and no other cancer code other than a Read code for the initial type of cancer has occurred in the interim

- If the initial cancer code is followed by two or more codes with Read term "Seen in oncology clinic" (9N1y800, 9N09.00) on different dates, and no other cancer code other than a Read code for the initial type of cancer has occurred in the interim, and no cancer care review code occurs on or before the date of the initial cancer

- If the index cancer is breast cancer, there are no Read codes for other types of cancer in the patient's records, no cancer care review code occurs on or before the date of the initial cancer, and the patient record includes one or more entries for prescriptions of common hormonal treatments after the index cancer code (i.e., any of the following terms show up anywhere in either the Product Name or Drug Substance Name fields:

aminoglutethimide, anastrozole, formestane, fulvestrant, goserelin, letrozole, tamoxifen, toremifene)

- If the index cancer is prostate cancer, there are no Read codes for other types of cancer in the patient's records, no cancer care review code occurs on or before the date of the initial cancer, and the patient record includes one or more entries for prescriptions of common hormonal treatments after the index cancer code (i.e., any of the following terms show up anywhere in either the Product Name or Drug Substance Name fields: aminoglutethimide, abiraterone, bicalutamide, enzalutamide, flutamide, goserelin, leuprorelin, nilutamide)

Free-Text Data Included in Medical Profile of Patients With Cancer

- We requested up to five free-text comments from clinical and referral records within 1 year following the initial cancer date.
- Priority was given to any events with a Read code for the initial cancer, in chronological order. If this resulted in fewer than five free-text comments being identified, then free text on additional records without Read codes for the initial cancer was selected in chronological order starting with the index date.
- If the patient had a second cancer at any point subsequent to the first cancer, we requested up to five additional free-text comments within 1 year on or following this second cancer code, using the same process to prioritize the free text around the time of the second cancer as was used for the initial cancer.
- Additionally, free text was requested on the first five records with a Read code for cancer care review and for all records with a code related to death.

Table: Cancer Cases Identifiable and Not Identifiable in General Practitioner Online Database Data During the Period of Complete Overlap of Data Sources (2004-2010) in Linked Practices, by Patient Characteristic

eTable. Cancer Cases Identifiable and Not Identifiable in the General Practitioner Online Database During the Period of Complete Overlap of Data Sources (2004-2010) in Linked Practices, by Patient Characteristic

	Overall				Pancreas				Lung and Bronchus				Kidney and Renal Pelvis			
	In GOLD data		Not in GOLD data		In GOLD data		Not in GOLD data		In GOLD data		Not in GOLD data		In GOLD data		Not in GOLD data	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
All	492	68%	228	32%	10	48%	11	52%	43	47%	48	53%	10	36%	18	64%
Age at outcome																
18-54	40	89%	5	11%	1	100%	0	0%	4	80%	1	20%	0	0%	1	100%
55-64	83	72%	32	28%	2	67%	1	33%	6	38%	10	63%	4	57%	3	43%
65-74	134	68%	64	32%	3	75%	1	25%	12	50%	12	50%	3	50%	3	50%
75+	235	65%	127	35%	4	31%	9	69%	21	46%	25	54%	3	21%	11	79%
Sex																
Male	230	69%	105	31%	0	0%	5	100%	24	56%	19	44%	7	44%	9	56%
Female	262	68%	123	32%	10	63%	6	38%	19	40%	29	60%	3	25%	9	75%
Year of outcome																
2004 and 2005	49	72%	19	28%	0	0%	2	100%	2	29%	5	71%	2	40%	3	60%
2006 to 2008	230	73%	87	27%	5	45%	6	55%	24	49%	25	51%	6	55%	5	45%
2009 and 2010	213	64%	122	36%	5	63%	3	38%	17	49%	18	51%	2	17%	10	83%
Exposure during follow-up																
Darifenacin	6	67%	3	33%	0	0%	0	0%	1	100%	0	0%	0	0%	0	0%
Fesoterodine	13	72%	5	28%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%
Oxybutynin	179	68%	86	32%	3	43%	4	57%	12	44%	15	56%	3	38%	5	63%
Solifenacin	109	68%	51	32%	1	50%	1	50%	8	42%	11	58%	4	44%	5	56%
Tolterodine	220	68%	105	32%	5	38%	8	62%	21	50%	21	50%	6	40%	9	60%
Trospium	123	67%	60	33%	3	75%	1	25%	13	43%	17	57%	2	50%	2	50%

	Overall				Pancreas				Lung and Bronchus				Kidney and Renal Pelvis			
	In GOLD data		Not in GOLD data		In GOLD data		Not in GOLD data		In GOLD data		Not in GOLD data		In GOLD data		Not in GOLD data	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Increased cardiovascular risk[a]	295	65%	156	35%	7	47%	8	53%	40	54%	34	46%	7	33%	14	67%
Index of Multiple Deprivation (quintiles)																
Q1	121	71%	50	29%	2	40%	3	60%	12	63%	7	37%	2	25%	6	75%
Q2	140	68%	66	32%	3	60%	2	40%	11	50%	11	50%	4	57%	3	43%
Q3	96	71%	40	29%	1	33%	2	67%	5	28%	13	72%	3	50%	3	50%
Q4	67	64%	37	36%	2	33%	4	67%	8	62%	5	38%	0	0%	3	100%
Q5	68	66%	35	34%	2	100%	0	0%	7	37%	12	63%	1	25%	3	75%
Smoking																
Current	59	61%	37	39%	1	33%	2	67%	16	52%	15	48%	1	17%	5	83%
Former	248	68%	119	32%	4	40%	6	60%	23	45%	28	55%	7	50%	7	50%
Nonsmoker	185	72%	71	28%	5	63%	3	38%	4	44%	5	56%	2	25%	6	75%
Unknown	0	0%	1	100%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%
Recorded overactive bladder diagnosis	280	70%	118	30%	5	50%	5	50%	23	44%	29	56%	4	33%	8	67%
Body mass index																
< 20	15	65%	8	35%	0	0%	0	0%	2	33%	4	67%	0	0%	0	0%
20 to < 25	113	73%	41	27%	4	57%	3	43%	10	53%	9	47%	0	0%	4	100%
25 to < 30	139	70%	59	30%	3	60%	2	40%	13	54%	11	46%	3	33%	6	67%
30 to < 40	81	65%	43	35%	1	50%	1	50%	4	29%	10	71%	3	60%	2	40%
40+	9	60%	6	40%	0	0%	0	0%	0	0%	1	100%	0	0%	1	100%
Unknown	135	66%	71	34%	2	29%	5	71%	14	52%	13	48%	4	44%	5	56%
Hypertension (diagnosis codes and/or medications)	467	68%	223	32%	10	48%	11	52%	41	47%	47	53%	10	37%	17	63%
Diabetes (diagnosis codes and/or medications)	74	62%	45	38%	5	63%	3	38%	13	59%	9	41%	2	29%	5	71%

GOLD = General Practitioner Online Database

[a] Increased cardiovascular risk was defined as having evidence or history of diabetes, acute myocardial infarction, heart failure, coronary heart disease, atrial fibrillation, stroke, transient ischemic attack, peripheral artery disease, or peripheral vascular disease; or meeting more than one of these criteria: being a current smoker, having a diagnosis of dyslipidemia, or having a diagnosis of hypertension.

Note: Cases not identifiable in the General Practitioner Online Database were those recorded in Hospital Episode Statistics and/or the National Cancer Data Repository but not in the General Practitioner Online Database. Patient characteristics were ascertained at the time of the cancer diagnosis. Shown in this table are the cancer types for which the General Practitioner Online Database seems to be least complete.

Similarities and Differences in Methods and Findings of Cancer Case Validation Studies in the UK

The validity of cancer diagnoses in the Clinical Practice Research Datalink (known as CPRD) has been shown to be high using a variety of validation methods.[3] Multiple studies have examined the completeness of cancer recording in data sources available for research in the UK. Dregan et al.[4] used the General Practitioner Online Database (primary care data within the Clinical Practice Research Datalink) without free-text comments plus cancer registry data from 2002-2006 to evaluate the concordance of diagnoses of lung, colorectal, gastroesophageal, and urinary tract cancers in 42,556 patients with signs or a diagnosis of these cancers.[4] In the General Practitioner Online Database, cancers were identified via diagnosis, morphology, and procedure codes; some benign neoplasms and in situ cancers were included. From a total of 5,923 cases in the General Practitioner Online Database or cancer registry data, the authors found more cancers in cancer registry data than in the General Practitioner Online Database for all cancer types. In our study, we also found more cases of these cancer types in the National Cancer Data Repository, but we found ascertainment in the General Practitioner Online Database to be less complete than reported by Dregan et al.[4]. For example, we found 36% of kidney cancers and 62% of bladder cancers in the Clinical Practice Research Datalink (in the period of overlap with the National Cancer Data Repository and Hospital Episodes Statistics; result not shown), compared with 86% of urinary tract cancers in Dregan et al.[4] One possible reason for the difference is that we would have missed cancer cases in General Practitioner Online Database that were not recorded using diagnosis codes. If neither Hospital Episode Statistics nor the National Cancer Data Repository were available, one could attempt to maximize case identification in the General Practitioner Online Database by screening for codes beyond diagnosis codes, but this would likely produce more false-positive cases because, for example, some morphology codes encompass both benign and malignant neoplasms.

Boggon et al.[5] used the General Practitioner Online Database and the National Cancer Data Repository from 1997-2006 to assess concordance of cancer diagnoses in patients aged at least 40 years with type 2 diabetes and matched patients without diabetes.[5] Free-text comments and Read codes for cancer therapy or visits to cancer clinics were used to confirm General Practitioner Online Database diagnoses found in the National Cancer Data Repository. Unlike Dregan et al.[4] and us, Boggon et al.[5] found more cases in the General Practitioner Online Database than in the National Cancer Data Repository. In the Boggon study, almost all cases in the National Cancer Data Repository were also identified in the General Practitioner Online Database (91%-99% for our 10 study cancers). In our study, results were lower and more variable, ranging from 42% (kidney cancer) to 92% (breast cancer); results not shown. Of note, Boggon et al.[5] reported that, of the National Cancer Data Repository cases, 11% were found in the General Practitioner Online Database only in free-text comments, and 0.3% only had therapy

codes (i.e., no diagnosis codes). These cases would not have been found in the General Practitioner Online Database in our study. Further, Boggon et al.[5] found that survival was better for patients with cancer identified in the General Practitioner Online Database than for patients with cancer identified in the National Cancer Data Repository. Death certificates are a key source for cancer registries in England, and it has been suggested that nonfatal cases may be underrecorded.[6] Additionally, and in agreement with our findings, Boggon et al.[5] and Rañopa et al.[7] reported that the General Practitioner Online Database was less complete for older patients than for younger ones.

Haynes et al.[8] speculated that incident colorectal cancer would be more completely recorded than other cancers in The Health Improvement Network (known as THIN) UK primary care database (which has a partial overlap with the Clinical Practice Research Datalink in terms of contributing practices), given the active role of general practitioners in screening; they did not find this to be the case.[8] However, a later study in The Health Improvement Network confirmed this hypothesis: Cea Soriano et al.[9] found that primary care data contained 94% of colorectal cancer cases identified in at least one of Hospital Episode Statistics or primary care data (including Read diagnosis codes, free-text comments, and some codes for referrals and personal history of cancer).[9] In our study, taking into consideration only the General Practitioner Online Database and Hospital Episode Statistics, only 70% of cases were identifiable in the General Practitioner Online Database (result not shown). Furthermore, an additional 7 cases (8% of all colorectal cancer cases in our study) were only in the National Cancer Data Repository and would have been missed had the National Cancer Data Repository not been used (result not shown).

Increased sensitivity in detecting cases is to be expected when combining data generated in the course of more than one type of health care encounter (e.g., primary care records from the General Practitioner Online Database and hospital data from Hospital Episode Statistics compared with the General Practitioner Online Database data only). Greater sensitivity may be obtained without a decrease in specificity if high proportions of cases from different sources are true cases. Ascertainment of cardiovascular endpoints has also been shown to be more complete when using multiple data sources within the Clinical Practice Research Datalink.[10, 11]

References

1. National Institute for Health and Clinical Excellence. Quality and outcomes framework (QOF) indicator development programme: indicator guidance. August 2012. https://www.nice.org.uk/Media/Default/standards-and-indicators/qof%20indicator%20key%20documents/NM62-QOF-Indicator-Guidance-Cancer.pdf Accessed August 18, 2015.

2. Department of Health. New BMS contract QOF implementation. Dataset and business rules—cancer indicator set. [Cancer ruleset v26.0. Version date: 01/06/2013]. June 1, 2013. http://cdn.pcc-cic.org.uk/sites/default/files/articles/attachments/cancer_ruleset_v26.0_0.pdf.pagespeed.ce.Y9Vy0rg2AU.pdf Accessed February 2, 2015.

3. 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;69:4-14.

4. 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;36:425-429.

5. 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;22:168-175.

6. Moller H, Richards S, Hanchett N, *et al.* Completeness of case ascertainment and survival time error in English cancer registries: impact on 1-year survival estimates. *Br J Cancer.* 2011;105:170-176.

7. Rañopa M, Douglas I, van Staa T, Smeeth L, Bhaskaran K. Validity of cancer diagnosis in UK primary care databases: comparison of observed and expected cancer incidence rates. *Pharmacoepidemiol Drug Saf* 2015;24:S532.

8. Haynes K, Forde KA, Schinnar R, Wong P, Strom BL, Lewis JD. Cancer incidence in The Health Improvement Network. *Pharmacoepidemiol Drug Saf.* 2009;18:730-736.

9. Cea Soriano L, Soriano-Gabarro M, Garcia Rodriguez LA. Validity and completeness of colorectal cancer diagnoses in a primary care database in the United Kingdom. *Pharmacoepidemiol Drug Saf.* 2016;25:385-391.

10. Herrett E, Shah AD, Boggon R, *et al.* Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *BMJ.* 2013;346:f2350.

11. Gallagher AM, Williams T, Leufkens HG, de Vries F. The impact of the choice of data source in record linkage studies estimating mortality in venous thromboembolism. *PLoS One.* 2016;11:e0148349.