PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2016; **25**: 385–391
Published online 5 October 2015 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/pds.3877

ORIGINAL REPORT

Validity and completeness of colorectal cancer diagnoses in a primary care database in the United Kingdom†

Lucía Cea Soriano[1], Montse Soriano-Gabarró[2] and Luis A. García Rodríguez[1]*

[1] *Spanish Centre for Pharmacoepidemiologic Research (CEIFE), Madrid, Spain*
[2] *Global Epidemiology, Bayer Pharma AG, Berlin, Germany*

ABSTRACT

Purpose To validate the recorded diagnoses of colorectal cancer (CRC) and identify false negatives in The Health Improvement Network (THIN) primary care database.

Methods We conducted a validation study of incident CRC cases in THIN among patients aged 40–89 years from 2000–2011. CRC Read code entries ($N = 3805$) were verified by manual review of patients' electronic medical records (EMRs) including free-text comments. Incident CRC cases in THIN ascertained following manual review were validated against two data sources deemed gold standards: (i) questionnaires sent to primary care practitioners (PCPs; for a random sample of 100 potential CRC cases), and (ii) Hospital Episode Statistics (HES) among linked practices. False negatives in THIN were identified by searching for International Classification of Diseases-10 codes related to CRC in HES.

Results Of 3805 CRC cases identified in THIN via Read codes, 3033 patients (80.0%) were considered definite cases after manual review of EMRs. The positive predictive value (PPV) of CRC Read codes was 86.0% after removing patients identified from THIN via a Read code for 'fast track referral for suspected CRC'. The response rate from PCPs was 87.0% ($n = 87$), and the PPV of CRC in THIN was 100% based on PCP questionnaires. Using HES, the PPV for CRC in THIN was 97.9% (556/568), and false negative rate was 6.1% (36/592).

Conclusions CRC diagnostic Read codes in THIN have a high PPV, which is increased further following manual review of free-text comments. The false negative rate of CRC diagnoses in THIN is low. © 2015 The Authors. *Pharmacoepidemiology and Drug Safety* published by John Wiley & Sons Ltd.

KEY WORDS—validation studies; colorectal cancer; database; pharmacoepidemiology

Received 22 June 2015; Revised 26 August 2015; Accepted 27 August 2015

INTRODUCTION

Colorectal cancer (CRC) is the third most common cancer in both males and females in the UK and the second most common cause of cancer death in the UK.[1] This study is part of a larger study designed to estimate the risk of CRC with use of low-dose aspirin in patients in the UK using data from The Health Improvement Network (THIN) primary care database.[2] THIN is one of several databases of electronic medical records (EMRs) arising from general practices throughout the UK, which are increasingly being used for pharmacoepidemiological research. They enable long-term follow-up of observational cohorts, and are able to provide large samples that are often representative of the target population. However, their utility in the evaluation of clinical outcomes is dependent on the validity of recorded diagnoses, and the extent to which cases of the outcome are captured.

Validation studies of a variety of medical conditions and outcomes in THIN have been undertaken previously, reporting high confirmation rates of recorded diagnoses,[3–11] yet the validity of CRC recording in THIN has yet to be established. In this study, we aimed to assess the validity of the recording of CRC diagnoses in THIN and identify false negatives in THIN. The study protocol was reviewed and approved by an independent scientific review committee (reference number 12-044V).

METHODS

Data source

THIN is a computerized database of anonymized electronic medical records (EMRs) comprising patient data

*Correspondence to: L. A. García Rodríguez, Spanish Centre for Pharmacoepidemiologic Research (CEIFE), Almirante 28, 28004 Madrid, Spain. E-mail: lagarcia@ceife.es

†This study was presented in poster format at the 31st International Conference on Pharmacoepidemiology & Therapeutic Risk Management, Boston, 2015.

that is systematically and prospectively recorded by primary care practitioners (PCPs) across the UK.[12] The database holds over 80 million patient years of patient data and covers approximately 6% of the UK population.[13] The computerized information includes clinical and administrative data which are entered by PCPs using Read codes or as free-text, and all prescriptions issued. Read codes are the standard clinical terminology used in UK general practice, supporting detailed clinical encoding of diagnoses, symptoms, laboratory tests and results, therapeutics, surgical procedures, and demographics.[14] Additional information obtained from hospital letters and emails can be entered retrospectively into the free-text section. PCPs may also maintain paper files with laboratory data, hospital discharge summaries, consultant letters, and other patient-specific information, which can be obtained by requesting copies of paper files and/or through surveys of PCPs without breach of confidentiality. For a subset of THIN practices, data can be linked at the patient level to Hospital Episode Statistics (HES)[15] (approximately 20% at the time of the study) HES contain clinical and administrative data on hospital episodes (admissions and visits), which are collected from UK National Health Service hospitals, and which are linked to International Classification of Diseases (ICD)-10 codes.

Study population

We evaluated the validity of CRC recording in THIN by establishing its positive predictive value (PPV) and completeness though a three-step process. Firstly, manual review of EMRs including free-text comments for patients with a CRC Read code entry. Incident CRC cases in THIN ascertained following manual review were then validated against two data sources deemed gold standards: (i) questionnaires sent to PCPs (for a random sample of 100 potential CRC cases) and (ii) HES among linked practices. Cases in this validation study came from part of a larger study that aimed to evaluate the association between risk of CRC and use of low-dose aspirin, and therefore comprise a subset of all CRC cases in THIN (Supplementary Figure 1). Briefly, cases were identified as having a first Read code for CRC (Supplementary Table S1) between January 2000 and December 2011 ($N = 3805$). They were required to be aged 40–89 years at diagnosis and have no record of cancer or prescription for low-dose aspirin prior to study entry.

Manual review of EMRs in THIN

The EMRs, including free-text comments, of all patients with a CRC Read code were manually reviewed. Patients were considered to be incident cases of CRC unless there was evidence from the medical records to indicate otherwise, e.g. no definite diagnosis following biopsy results, prevalent case, or where another primary cancer was present either concurrently or previously. Information relating to the CRC diagnosis was extracted, including (where available) details on site, stage, surgery, adjuvant therapy, and diagnostic procedures. The index date was the date of first symptom, screening or diagnostic procedure, or surgery, whichever came first. The index date was backdated from the CRC Read code date in the majority of cases (83%); the median number of backdated days was 36, and the mean was 56.6.

PCP questionnaires

Among the 3805 patients with a Read code for CRC, we selected a random sample of 100 (2.6%) patients, and a questionnaire was sent to the corresponding PCP. The questionnaire was designed to collect information about site of the CRC and whether the patient had undergone colonoscopy, and can be found in the Supplementary Methods and Materials. PCPs were also requested to confirm the CRC diagnosis and send copies of referral letters and other supporting information related to the diagnosis of CRC. Among patients for whom a completed questionnaire was returned, we calculated the PPV of the CRC diagnosis in THIN (ascertained following manual review) using the PCP-reported information as gold standard. We identified patients confirmed as incident CRC cases by both the PCP and following the THIN manual review process, and compared the information relating to the CRC diagnosis (e.g. site, stage, and treatment) from the two data sources. We restricted the comparison of each variable to cases with complete information for that variable from both data sources.

Linkage to HES

We used HES admission data as gold standard to calculate the following measures in THIN: PPV of the CRC diagnosis, proportion of false positives, and proportion of false negatives (CRC cases in HES not identified in THIN. HES data were available up to March 2011 and were considered gold standard based on the assumption that all cases of CRC were recorded unless patients attended a private clinic for surgery or adjuvant chemotherapy (estimated as 10–15% in England).

Validation of the CRC diagnosis in THIN using HES and false positives in THIN. Among all patients originally identified in THIN with a Read code suggestive of CRC ($N = 3805$), 728 were enrolled in practices

DOI: 10.1002/pds

linked to HES and had a CRC Read code date in THIN before 1 January 2011; this criterion was applied in order to have at least 3 months' data in HES after the diagnosis date in THIN. For these 728 patients, we identified those with a CRC ICD-10 code in HES (Supplementary Table S2) at any time and manually reviewed their HES records extracting all clinical information relating to the CRC diagnosis. Among patients classified as CRC cases in both THIN and HES following manual review of EMRs from both data sources ($N=509$), we compared the main clinical features of CRC between the two data sources.

Identification of false negatives in THIN using HES. Among members of the study population in THIN who were linked to HES but without a Read code for CRC ($N=64\,078$), we searched HES for patients with an ICD-10 code suggestive of CRC at any time. We discounted patients whose censoring date in THIN preceded the HES discharge date or was up to 30 days after (to account for possible delays in recording hospitalizations in THIN); patients with a record in HES for cancer other than CRC before the CRC hospitalization; and patients with a record of CRC in HES before their study entry date in THIN. These exclusion criteria were applied to identify only patients in HES who would have been at-risk of being detected as a CRC case in THIN. We calculated the number of false negatives in THIN by summing (i) additional CRC cases in HES (not detected in THIN) and (ii) CRC cases in HES that were classified as non-cases in THIN following manual review.

RESULTS

CRC cases in THIN

A total of 3033 of the 3805 potential computer-detected cases of CRC in THIN were classified as incident cases of CRC following the manual review process; a PPV of 79.7%. The site was colon in 61.9% of cases, rectum in 36.6%, and both in 1.5%. Information on CRC stage was available for 46.9% of cases. A total of 354 individuals were identified during follow-up with one of the two Read codes for 'fast track referral' rather than a diagnostic Read code, corresponding to 9.3% of all potential cases ($N=354/3805$). Of these, only 8.5% were confirmed as incident CRC cases providing a PPV for these two codes of less than 10%.

Among patients classified as non-cases ($n=772$), 294 (38.1%) were detected through a Read code for 'fast track referral suggestive of a possible CRC malignancy'. During the manual review process, none of these patients subsequently had a diagnosis of CRC recorded after being referred for investigation. If we had removed this Read code from the original code list used to identify CRC cases, the PPV would have been 86.4% (3033/3511). Also, among non-cases, 258 (33.4%) were excluded because they had a record of another primary cancer at or before the CRC diagnosis. Among these 258 patients, 118 (45.7%) could have been captured using a computer search for Read codes for other primary cancers during the study period and up to CRC diagnosis. The remaining 140 of these 258 cases were excluded based on information in the free-text comments during the manual review. Other reasons for exclusion are shown in Table 1. If we had not used the Read codes for 'fast track referral suggestive of a possible CRC malignancy' and 'Seen in fast track suspected colorectal cancer clinic' in the initial computer search, and had also removed patients identified by the computer search for another previous primary cancer, then a PPV of 89.4% (3033/3393) would have been obtained.

PCP questionnaires

Of the 100 questionnaires sent to PCPs, 87 were returned with complete information (87% valid response rate). The average age of these 87 patients (mean, 69.5 years; median 69.0 years) was similar to the average age of the 13 patients for whom the questionnaires returned did not contain complete information (mean, 70.4 years; median, 69.0 years). Of the 100 patients for whom PCP-information was sought, 80 had been classified as incident cases of CRC fol-

Table 1. Case classification after manual review of patient EMRs

CRC case classification	$N=3805$ *n* (%)
Case	3033 (80.0)
Non-case	772 (20.0)
Other primary cancer	258 (33.4)
Benign tumor*	24 (3.1)
Fast-track high-risk patient screening	294 (38.1)
Diagnosed before start date†	180 (23.3)
Updated THIN release‡	3 (0.4)
Non-confirmed	13 (1.7)

*Includes carcinoma in situ, benign polyp, and adenoma.
†Includes all patients identified any time before the study period by means of surgery, comments entered as free-text, or because of backdating the index date to the date of first symptom or diagnostic procedure.
‡We requested free-text comments using the latest available data from THIN at that time, whereas the computer search of CRC Read codes was undertaken with the previous available version of THIN. Upon review of the patient electronic records using the later available data, these previous entries had been removed.

lowing the THIN manual review process, and 20 had been classified as non-cases. Among the 87 questionnaires returned (71 patients were classified as cases and 16 as non-cases following manual review), 51 (58.6%) had additional documentation attached (e.g. letter from consultant, surgical procedures). PCPs confirmed the CRC diagnosis in all 71 patients deemed cases in THIN, and 14 of the 16 patients deemed non-cases in THIN (Table 2). For the two patients whom PCPs did not confirm non-case status, the PCP reported a diagnosis of CRC. During the THIN manual review, we had classified these patients as having a benign colorectal tumour.

The distribution of CRC stage, surgery, and adjuvant therapy was similar between THIN and the information provided by the PCP, while the distribution of site differed slightly between the two data sources (Table 3). There was a higher proportion of cases with CRC in the proximal colon when using data from the questionnaire compared with THIN (42.4% versus 37.9%). The location was the rectum in 31.8% based on the questionnaires and 40.9% in THIN, although it should be noted that in the THIN manual review, CRC situated in the rectosigmoid was classified as located in the rectum. These comparisons are all based on small absolute numbers and should be interpreted with caution.

Table 3. Features of CRC using information retrieved from THIN and PCP questionnaires among confirmed cases with information in both sources

	Confirmed CRC cases in THIN and by PCP ($N = 71$)	
	PCP questionnaire	THIN manual review*
Site	*66*	*66*
Colon proximal	28 (42.4)	25 (37.9)
Colon distal	17 (25.8)	14 (21.2)
Rectum†	21 (31.8)	27 (40.9)
Stage	*22*	*22*
Dukes A	6 (27.3)	6 (27.3)
Dukes B	5 (22.7)	6 (27.3)
Dukes C	10 (45.5)	7 (31.8)
Dukes D	1 (4.5)	3 (13.6)
Type of surgery	*39*	*39*
Hemicolectomy (left or right)	22 (56.4)	20 (51.3)
Abdominal perianal resection	4 (10.3)	3 (7.7)
Sigmoid colectomy	3 (7.7)	3 (7.7)
Hartmann's operation	2 (5.1)	—
Excised not specified	—	1 (2.6)
Anterior resection	6 (15.4)	10 (25.6)
Other	2 (5.1)	2 (5.1)

Data are *N* or *n* (%) as appropriate.
*Including review of free-text comments.
†CRC situated in the rectosigmoid was considered to be located in the rectum.
CRC, colorectal cancer; PCP, primary care practitioner; THIN, The Health Improvement Network.

PPV and false positives in THIN using HES

Of the 728 patients with a CRC Read code in THIN and linked to HES, 568 were classified as cases and 160 as non-cases in THIN following the manual review. Of the 568 incident CRC cases in THIN, 509 (89.6%) were also deemed to be incident cases in HES. Clinical features of CRC in these 509 patients are shown in Table 4. The CRC site was the colon in 57% of patients and the rectum in 43% of patients in both THIN and HES datasets. Surgical operations were found among 78.0% of CRC cases in HES and 73.9% of CRC cases in THIN, with hemicolectomy the most frequent surgery in both data sources. Adjuvant therapy was recorded in a greater proportion of cases in THIN (34.2%) than in HES (16.3%). When we restricted to CRC cases with complete information in both datasets for each variable analyzed, the distribution of CRC site and type of surgery was very similar between THIN and HES (Supplementary Table S3).Of the 568 CRC cases in THIN, 47 had no hospitalization because of CRC in HES (Figure 1), and 12 did not have their CRC diagnosis in THIN verified by HES data. Of these latter 12 patients, 11 were hospitalized for another primary cancer before the CRC diagnosis, and one patient was hospitalized before their THIN study entry date. These 12 patients were therefore misclassified as CRC cases in THIN, corresponding to a false positive rate of 2.1% (12/568). Subtracting these 12 patients from the 568 ascertained in THIN, corresponds to a PPV for CRC in THIN of 97.9% (556/568).

Table 2. Number of confirmed CRC cases in THIN and PPV using PCP questionnaires as gold standard

	Manual review of patient profiles in THIN*	
	Cases	Non-cases
Questionnaires sent to PCP	*N* (%; 95%CI)	*N* (%; 95%CI)
Total questionnaires sent	80	20
Valid questionnaires received	71 (88.8; 78.0–94.0)	16 (80.0; 58.4–91.9)
Confirmed case status	71 (100.0; 94.9–100.0)	14 (87.5; 64.0–96.5)
Non-confirmed case status	—	2 (12.5; 3.5–36.5)†

*Including free-text comments.
†These patients were considered to have a benign stage of carcinoma after manual review including the free-text comments.
CI, confidence interval.

Table 4. Characteristics of CRC cases in both HES and THIN

	CRC cases in both THIN and HES	
	Information retrieved in HES	Information retrieved in THIN
	N = 509	*N* = 509
	n (%)	*n* (%)
Site		
Colon proximal	145 (28.5)	140 (27.5)
Colon distal	116 (22.8)	107 (21.0)
Rectum	218 (42.8)	218 (42.8)
Colon unspecified	30 (5.9)	44 (8.6)
Surgery		
Yes	397 (78.0)	376 (73.9)
Not recorded/unknown	112 (22.0)	133 (26.1)
Type of surgery		
Hemicolectomy (left or right)	148 (37.3)	144 (38.3)
Abdominal perianal resection	41 (10.3)	26 (6.9)
Sigmoid colectomy	25 (6.3)	20 (5.3)
Hartmann's operation	27 (6.8)	16 (4.3)
Excised not specified	13 (3.3)	6 (1.6)
Anterior resection with/out anastomosis/colostomy	110 (27.7)	107 (28.5)
Panproctocolectomy	2 (0.5)	3 (0.8)
Transanal resection	—	2 (0.5)
Other	29 (7.3)	31 (8.2)
Unspecified	2 (0.5)	21 (5.6)
Adjuvant therapy		
Yes	83 (16.3)	174 (34.2)
Not recorded/unknown	426 (83.7)	335 (65.8%)

CRC, colorectal cancer; HES, Hospital Episode Statistics; THIN, The Health Improvement Network.

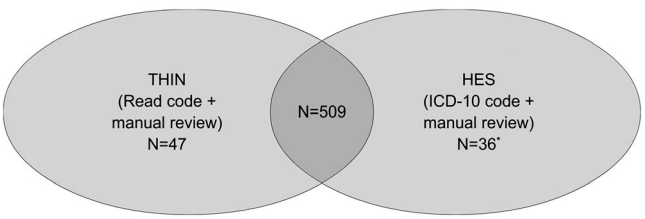

Figure 1. Concordance between CRC cases in THIN and HES. *Comprises 32 cases not captured in THIN plus four cases classed as non-cases in THIN following manual review (false negatives). HES, Hospital Episodes Statistics; ICD, International Classification of Diseases; THIN, The Health Improvement Network

False negatives in THIN using HES

Of the 160 patients classified as non-cases in THIN and linked to HES, four patients had a CRC diagnosis in HES that met the criteria for our operational definition of CRC. Among members of the study population in THIN who were linked to HES and without a Read code for CRC (*N* = 64,078), 506 patients had a CRC ICD-10 code in HES. After applying our exclusion criteria, 72 patients remained who were eligible to be, but were not, detected as a case of CRC in THIN. Of these, 40 had an ICD-10 code for 'personal history of malignant neoplasm of digestive organs' with no additional code for CRC, and therefore in the absence of additional information related to CRC were not considered to be CRC cases. Of the remaining 32 CRC cases in HES that were not identified in THIN, most (22, 68.8%) had records in THIN for diagnostic procedures, symptoms and/or specialist visits or had a discharge letter around the HES hospitalization date, yet did not have a definite CRC diagnosis recorded. Overall, considering there were 47 CRC cases ascertained only in THIN, 36 cases (32 + 4) only in HES and 509 cases in both THIN and HES (Figure 1), the corresponding false negative rate of CRC in THIN was 6.1% (36/592).

DISCUSSION

In this thorough validation of the recording of CRC in THIN, we have shown that automated computer searches for diagnostic CRC Read codes is a valid method for identifying incident cases of CRC in THIN, with a PPV of almost 90% when removing patients with a prior Read code for another primary cancer. However, Read codes for CRC fast track referral should not be included in such computer algorithms because of their low PPV. Furthermore, subsequent manual review of patients' EMRs increases the validity of using CRC diagnostic Read codes; PPVs were 100% using PCP-reported information as gold standard and 97.9% using HES. We also found the data in THIN regarding the clinical features of CRC to have a high level of consistency with the data provided by PCPs and HES. In line with previous studies in THIN, [3,8,10] our study highlights the value of the data entered as free-text. We found these data to be valuable not only in case identification, but also in obtaining additional clinical information relating to the diagnosis, such as cancer site and stage, and additional details relating to treatment, surgery and symptoms. We also found that some of the details obtained from the free-text review were not entered in HES; adjuvant therapy was recorded in twice as many patients in THIN as in HES.

Secondary care in the UK is predominantly accessed via PCP referral, with details on hospital visits and admissions communicated back to the PCP via letter or email, and updated in the primary care records retrospectively. The overall false negative rate in THIN was low at 6.1%, and of note is that the majority of cases in HES who were not ascertained as cases in THIN did have information recorded relating to diagnostics, symptoms or discharge letters, but no definite

recorded diagnosis. This indicates a high level of recording in THIN of the information obtained in secondary care.

The main strength of our study is the multi-step validation process, including large-scale manual review of patient's EMRs and validation using two data sources considered gold-standards. A high response rate (87.0%) was obtained for the PCP questionnaires, albeit a small sample size. We did not link to cancer registry data although this has been undertaken previously by others for 1992–2007.[16] Haynes *et al.* evaluated the recording of cancer diagnoses in both THIN and a UK national cancer registry, finding age- and sex-standardized incidence ratios for CRC to be close to unity in the latter years of their study period, particularly after 2004. Although this study did not validate CRC diagnoses in THIN, these findings support a high level of CRC recording in the database. In addition, a study using data from the UK Clinical Practice Research Datalink (CPRD), which contains similar primary care data to THIN, reported a 98% PPV for the CRC diagnosis in the primary care data when linked to cancer registrations.[17] A limitation inherent in some validation studies is that there is no true gold standard. In our study, 47 incident cases of CRC were identified in THIN that had no hospitalization relating to a diagnosis of CRC in HES, possibly because these patients attended a private hospital and therefore were not recorded in HES. The limitations of using various data sources in the UK as gold standard for a clinical diagnosis have been highlighted previously by others.[18] Another study reported an underestimation of incident CRC cases in CPRD primary care data when compared with registry data;[19] however, patients were required to have additional codes supporting the CRC diagnosis to be included as a case. We are aware of few other studies that have validated CRC diagnoses in other computerized healthcare databases. A study using a French administrative claims database reported PPVs of between 59% and 78% for the recording of new CRC cases compared with registry data, depending on the coding algorithm used.[20] In another study, Helqvist *et al.*[21] reported high quality ICD-10 CRC diagnosis coding data in the Danish National Registry of Patients using the Danish Cancer Registry as a reference, with a PPV of 89% and completeness rate of 93%.

Close to 400 research articles have been published using data from THIN,[13] including previous research on CRC.[2,22–25] The database has been shown to be representative of the UK population with regards to age, sex, and geographic distribution.[26] In addition, as part of the wider study from which this study arose,[2] we have found that the distribution of stage and site of the 3033 cases identified in THIN following manual review are broadly consistent with national data[27–29] supporting the representativeness to cases in the general population. Owing to its large size, THIN offers the potential to obtain precise risk estimates for clinical outcomes and provides information on important confounding variables and prescription data. Review of free-text comments can be a labour intensive process, especially for large cohorts, yet is essential when information relating to the clinical features of CRC (e.g. stage) are required to evaluate a particular research questions. For example, the effect of an exposure on the risk of CRC by stage at diagnosis, or the effect of a cancer treatment on survival according to CRC stage. However, for large-scale epidemiological studies involving CRC in THIN in which there is no necessity to obtain such clinical details (such as when CRC is included as a co-variate) use of diagnostic CRC Read codes is sufficient.

CONFLICT OF INTEREST

This work was supported by Bayer Pharma AG. Montse Soriano-Gabarró is a salaried, full-time employee of Bayer Pharma AG. Lucía Cea Soriano and Luis A. García Rodríguez work for CEIFE, which has received a research grant from Bayer Pharma AG. Dr García Rodríguez has also served as a consultant and advisory board member for Bayer Pharma AG. Bayer Pharma AG provided support in the form of salary for Montse Soriano-Gabarró, but had no role in the study design, the collection, analysis, and interpretation of data, nor in the writing of the report nor the decision to submit the report for publication.

KEY POINTS

- THIN is a valid resource for conducting large-scale epidemiologic studies of CRC using Read codes.
- CRC diagnoses in THIN had high PPVs and a low false negative rate following thorough review of clinical information, including free-text comments.
- For CRC outcome studies in THIN that require information on the clinical features to answer the research question, review of free-text comments is essential.

ETHICS STATEMENT

The study protocol was reviewed and approved by an independent scientific review committee (reference number 12-044V).

ACKNOWLEDGEMENTS

We thank Susan Bromley, EpiMed Communications Ltd (Oxford, UK) for help with the manual review of patient records and for providing medical writing assistance funded by Bayer Pharma AG. We are also grateful to Angel Lanas and Matias Cea Soriano for their contribution and helpful comments on the review of CRC cases.

REFERENCES

1. Cancer Research UK. Bowel Cancer Incidence Statistics. http://www.cancerresearchuk.org/cancer-info/cancerstats/types/bowel/incidence/ [23 December 2014].
2. García Rodríguez LA, Soriano-Gabarró M, Cea Soriano L. Low-dose acetylsalicylic acid and risk of colorectal cancer: evidence from a primary care study in the UK. *Gastroenterology* 2015; **148**(4Supplement 1): S-371.
3. Gaist D, Wallander MA, Gonzalez-Perez A, *et al.* Incidence of hemorrhagic stroke in the general population: validation of data from The Health Improvement Network. *Pharmacoepidemiol Drug Saf* 2013; **22**(2): 176–182.
4. Lo Re V, 3rd, Haynes K, Forde KA, *et al.* Validity of The Health Improvement Network (THIN) for epidemiologic studies of hepatitis C virus infection. *Pharmacoepidemiol Drug Saf* 2009; **18**(9): 807–814.
5. Ogdie A, Alehashemi S, Love TJ, *et al.* Validity of psoriatic arthritis and capture of disease modifying antirheumatic drugs in the health improvement network. *Pharmacoepidemiol Drug Saf* 2014; **23**(9): 918–922.
6. Mamtani R, Haynes K, Boursi B, *et al.* Validation of a coding algorithm to identify bladder cancer and distinguish stage in an electronic medical records database. *Cancer Epidemiol Biomarkers Prev* 2015; **24**(1): 303–307.
7. Margulis AV, Garcia Rodriguez LA, Hernandez-Diaz S. Positive predictive value of computerized medical records for uncomplicated and complicated upper gastrointestinal ulcer. *Pharmacoepidemiol Drug Saf* 2009; **18**(10): 900–909.
8. Martin-Merino E, Fortuny J, Rivero E, *et al.* Validation of diabetic retinopathy and maculopathy diagnoses recorded in a U.K. primary care database. *Diabetes Care* 2012; **35**(4): 762–767.
9. Meal A, Leonardi-Bee J, Smith C, *et al.* Validation of THIN data for non-melanoma skin cancer. *Qual Prim Care* 2008; **16**(1): 49–52.
10. Ruigómez A, Martin-Merino E, Rodriguez LA. Validation of ischemic cerebrovascular diagnoses in the health improvement network (THIN). *Pharmacoepidemiol Drug Saf* 2010; **19**(6): 579–585.
11. Seminara NM, Abuabara K, Shin DB, *et al.* Validity of The Health Improvement Network (THIN) for the study of psoriasis. *Br J Dermatol* 2011; **164**(3): 602–609.
12. Bourke A, Dattani H, Robinson M. Feasibility study and methodology to create a quality-evaluated database of primary care data. *Inform Prim Care* 2004; **12**(3): 171–177.
13. CSD Medical Research UK. http://csdmruk.cegedim.com/[6 January 2015].
14. Stuart-Buttle CD, Read JD, Sanderson HF, *et al.* A language of health in action: Read Codes, classifications and groupings. *Proc AMIA Annu Fall Symp* 1996: 75–79.
15. Health & Social Care Information Centre. Hospital Episode Statistics. http://www.hscic.gov.uk/hes [02/01/2015].
16. Haynes K, Forde KA, Schinnar R, *et al.* Cancer incidence in The Health Improvement Network. *Pharmacoepidemiol Drug Saf* 2009; **18**(8): 730–736.
17. Dregan A, Moller H, Murray-Thomas T, *et al.* 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): 425–429.
18. 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).
19. Charlton R, Snowball J, Bloomfield K, *et al.* Colorectal cancer incidence on the General Practice Research Database. *Pharmacoepidemiol Drug Saf* 2012; **21**(7): 775–783.
20. Quantin C, Benzenine E, Hagi M, *et al.* Estimation of national colorectal-cancer incidence using claims databases. *J Cancer Epidemiol* 2012; **2012**(298369).
21. Helqvist L, Erichsen R, Gammelager H, *et al.* Quality of ICD-10 colorectal cancer diagnosis codes in the Danish National Registry of Patients. *Eur J Cancer Care (Engl)* 2012; **21**(6): 722–727.
22. Boursi B, Haynes K, Mamtani R, *et al.* Digoxin use and the risk for colorectal cancer. *Pharmacoepidemiol Drug Saf* 2014; **23**(11): 1147–1153.
23. Damery S, Ryan R, Wilson S, *et al.* Iron deficiency anaemia and delayed diagnosis of colorectal cancer: a retrospective cohort study. *Colorectal Dis* 2011; **13**(4): e53–e60.
24. Hamilton W, Lancashire R, Sharp D, *et al.* The importance of anaemia in diagnosing colorectal cancer: a case–control study using electronic primary care records. *Br J Cancer* 2008; **98**(2): 323–327.
25. Osborn DP, Limburg H, Walters K, *et al.* Relative incidence of common cancers in people with severe mental illness. Cohort study in the United Kingdom THIN primary care database. *Schizophr Res* 2013; **143**(1): 44–49.
26. Blak BT, Thompson M, Dattani H, *et al.* Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. *Inform Prim Care* 2011; **19**(4): 251–255.
27. McPhail S, Johnson S, Greenberg D, *et al.* Stage at diagnosis and early mortality from cancer in England. *Br J Cancer* 2015; **112**(Suppl 1): S108–S115.
28. Public Health England. National Cancer Intelligence Network. Colorectal Cancer Survival by Stage—NCIN Data Briefing. http://www.ncin.org.uk/publications/data_briefings/colorectal_cancer_survival_by_stage [25 November 2014].
29. Association of Coloproctology of Great Britain and Ireland/The Northern and Yorkshire Cancer Registry and Information Service/The NHS Information Centre for Health and Social Care. The National Bowel Cancer Audit Annual Report 2009.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web site.