OPEN

Epidemiology Publish Ahead of Print

DOI: 10.1097/EDE.0000000000000786

Validation of Cancer Cases Using Primary Care, Cancer Registry, and Hospitalization Data in the UK

Andrea V Margulis [a,] Joan Fortuny [a,] James A Kaye [b,] Brian Calingaert [c,] Maria Reynolds [c], Estel Plana [a,] Lisa J McQuay [c,] Willem Jan Atsma [d,] Billy Franks [d,] Stefan de Vogel [d,] Susana Perez-Gutthann [a], Alejandro Arana [a]

[a] RTI Health Solutions, Barcelona, Av. Diagonal, 605, 9-1, 08028 Barcelona, Spain

[b] RTI Health Solutions, Waltham, 307 Waverley Oaks Road, Suite 101, Waltham, MA 02452, USA

[c] RTI Health Solutions, RTP, 200 Park Offices Drive, Research Triangle Park, NC 27709, USA

[d] Astellas Pharma B.V., PO Box 344, 2300 AH, Leiden, The Netherlands

Corresponding Author: Andrea V Margulis, RTI Health Solutions, Av. Diagonal, 605, 9-1, 08028 Barcelona, Spain , Telephone: +34.93.241.7766, E-mail: amargulis@rti.org

Running head: Validation of common cancers in UK primary care

Financial Support: This study was funded by Astellas Pharma Global Development, Inc.

Conflicts of Interest: The contract provides the research team independent publication rights. Andrea Margulis, Joan Fortuny, James Kaye, Brian Calingaert, Maria Reynolds, Estel Plana, Lisa McQuay, Susana Perez-Gutthann, and Alejandro Arana are employees of RTI International, an independent, nonprofit research organization that does work for government

agencies and pharmaceutical companies. Willem Jan Atsma, Billy Franks, and Stefan de Vogel are employees of Astellas Pharma Global Development, the sponsors of this study.

Acknowledgments: We thank Jennifer Bartsch for her help with programming, Christine Bui and Alicia Gilsenan for their help managing the project, Adele Monroe for her editorial help, Jason Mathes for his help preparing figures (all from RTI International); Kwame Appenteng and Milbhor D´Silva for their input at all stages of the study (both from Astellas); and CPRD research staff for their support to the program.

Reproducibility: The results of the study were generated by RTI Health Solutions (RTI-HS) using data obtained from CPRD. RTI-HS developed proprietary code to perform the analyses on the data. Researchers desiring access to the data would be required to obtain permission from the study sponsor, obtain data use agreement with CPRD and develop their own code.

Word count for abstract: 248 (including headings)

Word count for main text (excluding references): 2,187 (3,743 including (a) main text (for research articles, this typically includes the introduction, methods, results, and discussion), (b) bibliography, (c) tables, (d) figure legends, and (e) figures, calculated as 250 words per figure)

Total number of pages: 18 (abstract, registration statement, body of manuscript, references, 1 table and 3 figures)

Number of text pages: 13 (abstract, registration statement, body of manuscript, references)

Number of table pages: 1

Number of figure pages: 3

ABSTRACT

Background: In the United Kingdom, hospital or cancer registry data can be linked to electronic medical records for a subset of general practices and years.

Methods: We used Clinical Practice Research Datalink data (2004-2012) from patients treated for overactive bladder. We electronically identified provisional cases of 10 common cancers in General Practitioner Online Database data and validated them by medical profile review. In practices with linkage to Hospital Episodes Statistics and National Cancer Data Repository (2004-2010), we validated provisional cancer cases against these data sources. This linkage also let us identify additional cancer diagnoses in individuals without cancer diagnosis records in the General Practitioner Online Database.

Results: Among 50,840 patients, 1,486 provisional cancer cases were identified in the General Practitioner Online Database for 2004-2012. Medical profile review confirmed 93% of 661 cases in non-linked practices (range, 100% of non-Hodgkin lymphomas and uterine cancer to 77% of skin melanomas) and 96% of 825 cases in linked practices (100% of kidney and uterine cancers to 92% of melanomas). In the subset of linked practices, for 2004-2010, 720 cases were confirmed, of which 68% were identifiable in the General Practitioner Online Database (range, 90% of breast to 36% of kidney cancers).

Conclusions: Most cases of cancer identified electronically in the General Practitioner Online Database were confirmed. A substantial proportion of cases, especially of cancer types not typically managed by general practitioners, would be missed without Hospital Episodes Statistics and National Cancer Data Repository data (and are likely missed in non-linked practices).

Registration (before study conduct): European Union electronic Register of Post-Authorisation Studies **(**EU PAS Registry) number EUPAS5529, http://www.encepp.eu/encepp/viewResource.htm?id=11107

Keywords: Neoplasms; Validation studies; Electronic health records; Hospital Records; Registries; United Kingdom; CPRD

INTRODUCTION

Electronic medical records generated during routine primary care in the United Kingdom (UK) are often used for health care research. The capture of cancer cases in primary care electronic medical records, such as the General Practitioner Online Database, the primary care part of the Clinical Practice Research Datalink (known as CPRD), has been shown to be incomplete and to vary by cancer type.[1-3]

To increase validity and completeness, use of additional data sources may be warranted, such as hospital records (Hospital Episodes Statistics) or cancer registry data (National Cancer Data Repository), but these data sources are available only for a subset of patients in the Clinical Practice Research Datalink and are not available for the most recent patient follow-up due to data lag (about 1 year for Hospital Episode Statistics and 2 years for the National Cancer Data Repository).

As a part of an international postapproval cancer safety program evaluating a new drug to treat overactive bladder, we validated cancer endpoints in the General Practitioner Online Database and linked data to Hospital Episode Statistics and the National Cancer Data Repository.[4] The results from this validation effort are presented here.

METHODS

Data sources

The Clinical Practice Research Datalink, covering about 7% of the UK population, contains electronic medical records created by general practitioners during their clinical practice. General practitioners provide referrals to specialists, receive results from specialists and hospital discharge notes, and prescribe treatment for acute and chronic conditions.[5] The General Practitioner Online Database includes issued prescriptions and Read codes for

diagnoses, signs, symptoms, referrals, test requests, and test results, as well as free-text comments, which are unstructured fields for information supplementing coded entries. Information is recorded to the extent that it is important for health care. About 75% of English practices contributing to the Clinical Practice Research Datalink have consented to have their patients' information linked to other health care data sets, like Hospital Episode Statistics or the National Cancer Data Repository.[5] In Hospital Episode Statistics and the National Cancer Data Repository, diagnoses are recorded using the *International Statistical Classification of Diseases and Related Health Problems, 10th Revision* (ICD-10). All data for this study were de-identified.

In the parent cancer safety study, patients with a prescription for darifenacin, fesoterodine, oxybutynin, solifenacin, tolterodine, or trospium were included if they had at least 12 months of continuous enrollment before the prescription in an "up-to-standard" practice (a practice considered by the Clinical Practice Research Datalink to deliver data of adequate quality for research), provided that the same agent was not prescribed during the previous 12 months and that the patient was 18 years or older at the time of the prescription. We excluded patients with previous cancer (except non-melanoma skin cancer) because the focus of this study was first incident cancers. Patients with human immunodeficiency (HIV) infection were excluded because these patients may have received health care through specialty clinics or separate health plans, and their health service utilization might not be captured fully in the Clinical Practice Research Datalink.

Validation cohort

For the validation study, we selected from the population included in the safety study a stratified random sample, retaining all patients with a qualifying prescription for the three

least commonly prescribed drugs (darifenacin, fesoterodine, and trospium) and 33% of patients with a qualifying prescription for the most common drugs (oxybutynin, solifenacin, and tolterodine). This was done to ensure that all study drugs would be well represented in the validation cohort.

The study period was 1 January 2004 to 31 December 2012. Because the end of data collection in the General Practitioner Online Database was later than in Hospital Episode Statistics and the National Cancer Data Repository, linked person-time in patients enrolled in practices with linkage to Hospital Episode Statistics and the National Cancer Data Repository was followed by non-linked person-time. The period of complete overlap between data sources was 1 January 2004 to 31 December 2010 (Figure 1). Follow-up started with the qualifying prescription and ended at the earliest of end of the study period, disenrollment, HIV infection or cancer (except non–melanoma skin cancer), or death. We conducted validation efforts on the validation cohort.

The cancer endpoints were 10 common cancers: bladder, female breast, colorectal, corpus uteri, kidney and renal pelvis, lung and bronchus, non-Hodgkin lymphoma, pancreas, prostate, and skin melanoma.

Case identification and validation

Validation processes available for each patient depended on whether the individual's data in the General Practitioner Online Database were linked to Hospital Episode Statistics and the National Cancer Data Repository.

Validation in the General Practitioner Online Database (non-linked and linked) practices

Provisional cancer cases were identified using an electronic algorithm that searched for Read diagnosis codes in the General Practitioner Online Database, for practices without or with

linkage to Hospital Episode Statistics and the National Cancer Data Repository. As morphology and treatment codes are often not specific to cancer type, we did not include these types of codes in the electronic algorithm; we used them for case confirmation. Codes for benign neoplasms and in situ cancers were not included in the electronic algorithm.

We created electronic medical profiles with patients' diagnoses, procedures, relevant additional clinical information, and prescriptions. Medical profiles for these patients were reviewed by a team of clinical reviewers blinded to the study drugs, including a specialist in medical oncology/hematology, with free-text comments around the event date (n=405; free-text comments were requested when the diagnosis was not clear from the cancer-related codes) or without free-text comments (n=1,081). Provisional cases identified by the electronic algorithm were confirmed when patient medical profiles presented supportive clinical evidence of a cancer diagnosis, including morphology and treatment codes, codes indicating the general practitioner reviewed the patient´s cancer care, or supportive free-text comments. Details on the content of patient profiles, criteria to request free-text comments, reviewers' training, and review process are presented in the supplemental information.

If definitive information was found indicating that a provisional case did not have a cancer diagnosis, the patient was considered a non-case. When the medical profile had evidence that a provisional case had cancer diagnosed before cohort entry, the patient was considered a non-case and excluded from the study. Provisional cases not confirmed and not identified as non-cases remained provisional. Reviewers also assessed cancer type and diagnosis date. Discrepancies or uncertainties were reviewed by the team and resolved by the clinical specialist in medical oncology/hematology (JAK). The diagnosis date was the earliest date of a cancer diagnosis in any of the sources.

Additional validation in linked practices

For linked practices, validation started with the identification of provisional cases using the electronic algorithm previously described, followed by physician review of medical profiles. During the period of overlap between data sources (2004-2010), we used Hospital Episode Statistics and the National Cancer Data Repository to confirm previously identified cases (patients with cancer records in the General Practitioner Online Database and cancer records in one or both of these linked data sources) and to identify additional cases (patients in the General Practitioner Online Database without cancer records in the General Practitioner Online Database but with cancer records in Hospital Episode Statistics and/or the National Cancer Data Repository). Since Hospital Episode Statistics data are independently audited and cancer registries perform their own independent case validation using standardized procedures, including review of pathology information,[6] all cases identified in Hospital Episode Statistics or the National Cancer Data Repository were considered confirmed.

Statistical analysis

Based only on the General Practitioner Online Database from cases for the entire period, we reported the number of cancer cases identified using an electronic algorithm, plus absolute and relative frequencies of case confirmation from electronic medical profile review, overall and by linkage availability. We reported the frequency of cancer cases identifiable and not identifiable in the General Practitioner Online Database from linked practices, within the period with complete overlap of data sources, by patient characteristics, for selected cancer types. We described the source of each confirmed cancer diagnosis (General Practitioner Online Database, Hospital Episode Statistics, and/or the National Cancer Data Repository) using proportional Venn diagrams for the combined study cancers and for individual cancer

types. The area of each segment in these diagrams is proportional to the number of patients it includes.

Analyses were conducted using SAS 9.3 (Cary, NC: SAS Institute, Inc.; 2011) and Stata 13.1 (College Station, TX: StataCorp LP; 2014). The study protocol was registered in the European Union electronic Register of Post-Authorisation Studies before the study was conducted (Register number EUPAS5529; http://www.encepp.eu/encepp/viewResource.htm?id=11107) and was approved by the Clinical Practice Research Datalink's Independent Scientific Advisory Committee (protocol 13_142A).

RESULTS

Participants

The validation cohort included 50,840 study drug users. After excluding patients with cancer or HIV before cohort entry, the electronic search identified 1,486 provisional cancer cases in the General Practitioner Online Database, 56% from linked and 44% from non-linked practices.

Validation of provisional cases using only the General Practitioner Online Database, entire study period

Of the 1,486 provisional cancer cases identified through an electronic algorithm in the General Practitioner Online Database, 95% were confirmed in the review of patient's medical profiles (Table). Of the 825 provisional cases from linked practices, 96% were confirmed; at least 90% of provisional cases were confirmed for any individual cancer type. Of the 661 provisional cases in non-linked practices, 93% were confirmed in the review of medical patient profiles. For most individual cancer types (i.e., bladder, breast, colorectal,

corpus uteri, non-Hodgkin lymphoma, pancreas, and prostate), at least 90% of provisional cases were confirmed; for lung and kidney cancer and skin melanoma, 77% to 88% of provisional cases were confirmed.

Source of cases in linked practices using all data sources, period of overlap

Overall, 720 cancer cases were confirmed in the General Practitioner Online Database, Hospital Episode Statistics, and/or the National Cancer Data Repository. Of these, 68% were identifiable in the General Practitioner Online Database, 81% in Hospital Episode Statistics, and 84% in the National Cancer Data Repository (Figure 2). The completeness of case recording in the General Practitioner Online Database was greater for breast cancer and prostate cancer than for other study cancers (Figure 3).

In the General Practitioner Online Database, more complete identification of study cancer cases was seen in younger individuals (eTable; http://links.lww.com/EDE/B297 in Supplemental Digital Content), in non-smokers, and in cancers diagnosed in 2004-2008. Based on other characteristics, no substantial variation was apparent for the combined study cancers or for three cancers for which the General Practitioner Online Database is less complete: pancreas, lung, and kidney.

DISCUSSION

A very high proportion of provisional cases of cancer identified in the General Practitioner Online Database by screening for Read diagnosis codes were confirmed through clinical review of patient profiles or linkage to the National Cancer Data Repository or Hospital Episode Statistics, but, of these three data sources, no single source contained records of all confirmed study cancer cases. Completeness of cancer recording in the General Practitioner Online Database is higher for breast and prostate cancers—diseases for which general

practitioners often prescribe ongoing drug therapy—than for other cancers that are usually treated by specialists. We observed more complete case ascertainment in younger individuals, but we did not identify patient groups for which the General Practitioner Online Database contains all cancer cases.

Multiple studies have examined the completeness of cancer recording in data sources available for research in the UK. A discussion of methods and findings in our and other studies is presented in the supplemental information.

Cancer ascertainment from practices whose data allow linkage to Hospital Episode Statistics and the National Cancer Data Repository is more complete than from non-linked practices. Whether this would affect relative risks in safety studies depends on whether completeness of case ascertainment is differential for patients with versus without the exposure of interest. Even without such bias, a lower proportion of identified cases of a given cancer would be expected to yield more imprecise effect estimates.

A limitation of this study is that we identified cancer cases in the General Practitioner Online Database with an algorithm that used Read diagnosis codes exclusively (i.e., without morphology or treatment codes). While the coding system includes codes for morphology (e.g., Read code BB5..11, "[M] Adenocarcinoma") and treatment (e.g., Read code 8BAD.00, "Chemotherapy"), only diagnosis codes consistently permit identification of the type of cancer (needed in this study). Instead, in medical profile review, morphology and treatment codes were used, along with codes related to review of cancer care, to confirm the presence of cancer. Strengths of this validation study include the meticulous process for patient profile review and confirmation of case status, including calibration of the assessment process before starting the patient profile review to decrease inter-rater variability.

In conclusion, cancer case identification in the General Practitioner Online Database is sensitive to features of the case ascertainment algorithm such as the use of free-text comments and the type of codes included (e.g., diagnosis, morphology, treatment). Nearly all cancers with diagnosis codes in the General Practitioner Online Database (similarly for linked and non-linked practices) were confirmed. While completeness of the General Practitioner Online Database was high for breast and prostate cancer, a substantial proportion of other cancers will be missed if Hospital Episode Statistics and the National Cancer Data Repository are not used.

REFERENCES

1. 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.
2. 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.
3. Rañopa M, Douglas I, van Staa T, *et al.* The identification of incident cancers in UK primary care databases: a systematic review. *Pharmacoepidemiol Drug Saf.* 2015;24:11-18.
4. Kaye JA, Margulis AV, Fortuny J, *et al.* Cancer incidence after initiation of antimuscarinic medications for overactive bladder in the United Kingdom: evidence for protopathic bias. *Pharmacotherapy.* 2017: doi: 10.1002/phar.1932. [Epub ahead of print].
5. Herrett E, Gallagher AM, Bhaskaran K, *et al.* Data resource profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol.* 2015;44:827-836.
6. Jensen OM, Parkin DM, MacLennan R, Muir CS, Skeet RG, editors. *Cancer registration: principles and methods. IARC Publication No. 95*. Lyon, France: World Health Organization, International Agency for Research on Cancer (IARC), and International Association of Cancer Registries; 1991.

FIGURE LEGENDS (FOOTNOTES)

Footnote to figure 1

GOLD = General Practitioner Online Database; HES = Hospital Episode Statistics; NCDR = National Cancer Data Repository; ONS = Office for National Statistics.

Footnote to figure 2

GOLD = General Practitioner Online Database; HES = Hospital Episode Statistics; NCDR = National Cancer Data Repository.

Note: This figure represents the 720 confirmed cases in linked practices, regardless of the data source in which the cases were initially identified. Percentages were calculated using 720 as the denominator.

Footnote to figure 3

GOLD = General Practitioner Online Database.

Note: Each set of three circles represents confirmed cancer cases found in the General Practitioner Online Database (orange circle) and in Hospital Episode Statistics and the National Cancer Data Repository (black circles). Areas are proportional to the number of cases found in each source.

Table. Results of Validation of Provisional Cases Based Only on the General Practitioner Online Database in the Entire Study Period (2004-2012) in Linked and Non-Linked Practices

	All Practices			Linked Practices			Non-linked Practices		
	Identified in GOLD With Electronic Algorithm	Confirmed in Review of Medical Profile		Identified in GOLD With Electronic Algorithm	Confirmed in Review of Medical Profile		Identified in GOLD With Electronic Algorithm	Confirmed in Review of Medical Profile	
	n	n	%	N	n	%	n	n	%
Any study cancer	1,486	1,408	95%	825	792	96%	661	616	93%
Bladder[a]	179	170	95%	92	89	97%	87	81	93%
Breast	361	355	98%	208	205	99%	153	150	98%
Colorectal	198	187	94%	106	102	96%	92	85	92%
Corpus uteri	44	44	100%	27	27	100%	17	17	100%
Kidney and renal pelvis	31	29	94%	15	15	100%	16	14	88%
Lung and bronchus	165	149	90%	87	81	93%	78	68	87%
Non-Hodgkin lymphoma	47	46	98%	32	31	97%	15	15	100%
Pancreas	45	43	96%	25	24	96%	20	19	95%
Prostate[a]	344	325	94%	196	185	94%	148	140	95%
Skin melanoma	71	60	85%	36	33	92%	35	27	77%

GOLD = General Practitioner Online Database

[a] One patient had codes for bladder and prostate cancer on the same day.

Figure 1. Data Source Coverage in Relation to the Study Period

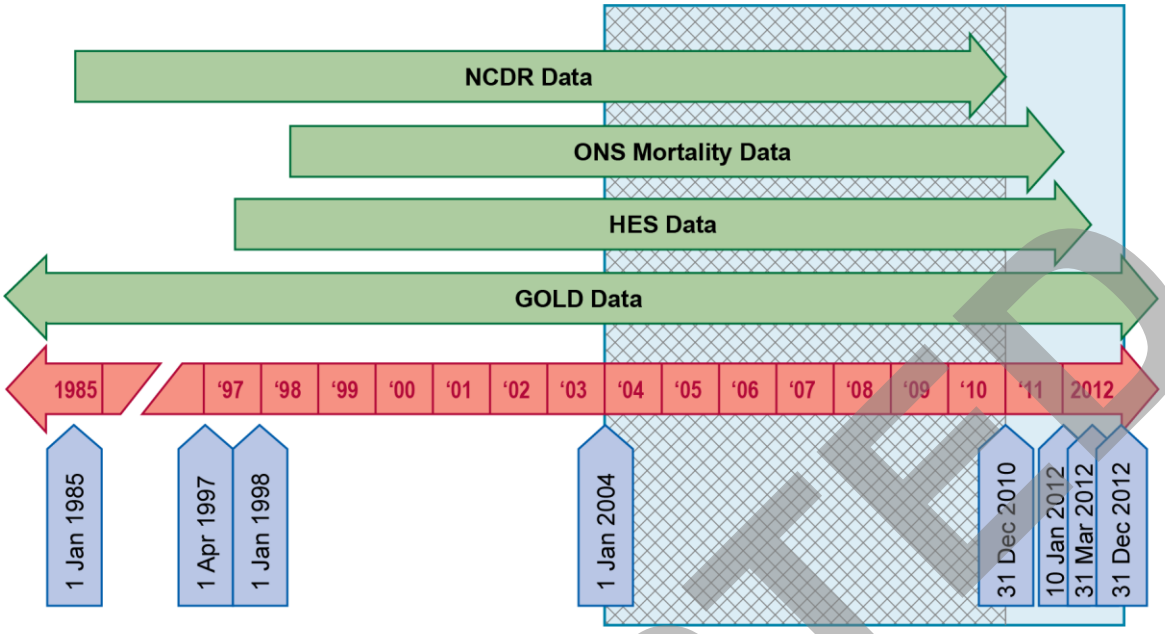

Figure 2. Origin of Cancer Cases Diagnosed During Period of Complete Overlap of Data Sources (2004-2010) in Linked Practices, by Data Source All Study Cancers Combined

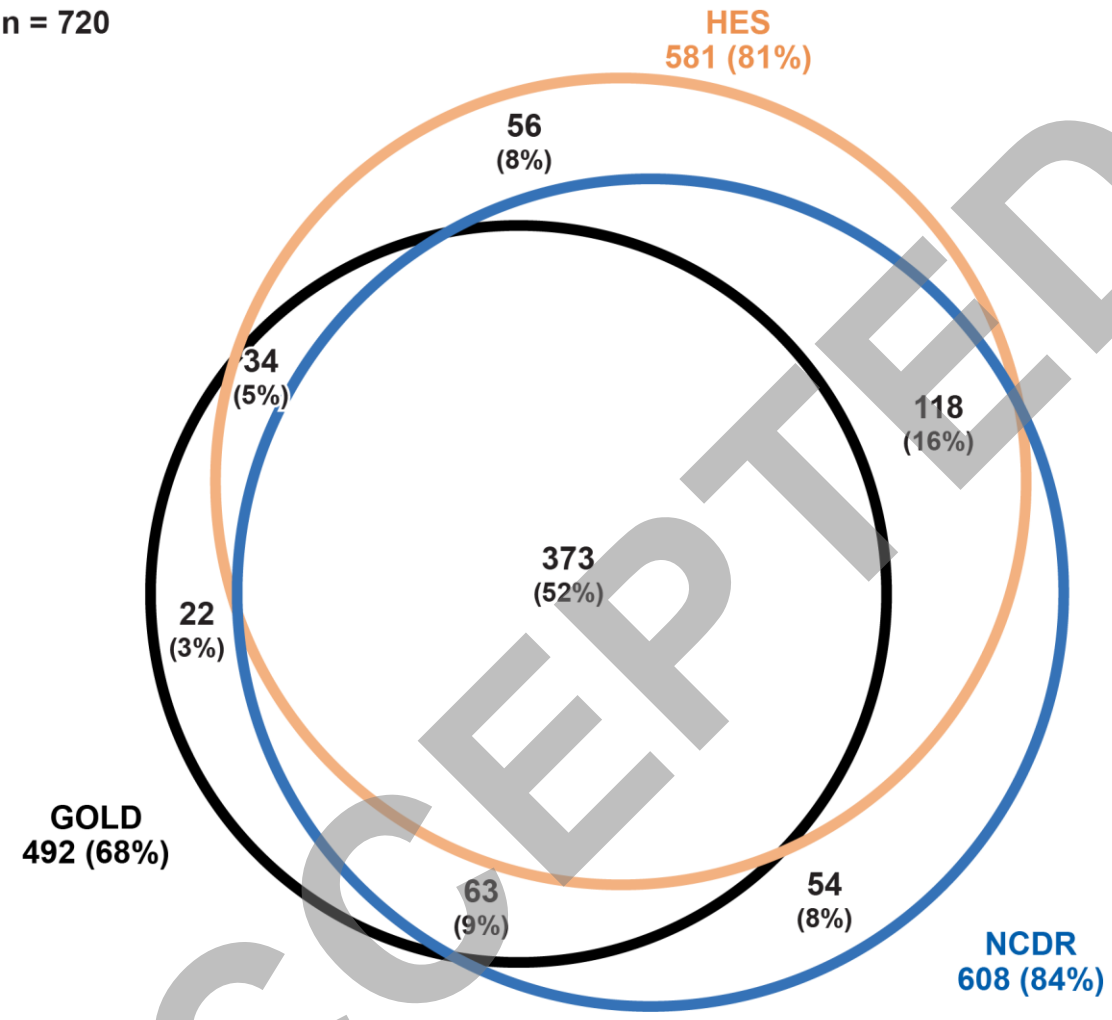

Figure 3. Selected Cancers by Main Treating Physician: Percentage of Cases Identifiable in the General Practitioner Online Database During the Period of Complete Overlap of Data Sources (2004-2010) in Linked Practices