Drug Saf (2013) 36:119–134
DOI 10.1007/s40264-012-0009-3

ORIGINAL RESEARCH ARTICLE

An Evaluation of the THIN Database in the OMOP Common Data Model for Active Drug Safety Surveillance

Xiaofeng Zhou · Sundaresan Murugesan · Harshvinder Bhullar · Qing Liu · Bing Cai · Chuck Wentworth · Andrew Bate

Published online: 4 January 2013

Abstract

Background There has been increased interest in using multiple observational databases to understand the safety profile of medical products during the postmarketing period. However, it is challenging to perform analyses across these heterogeneous data sources. The Observational Medical Outcome Partnership (OMOP) provides a Common Data Model (CDM) for organizing and standardizing databases. OMOP's work with the CDM has primarily focused on US databases. As a participant in the OMOP Extended Consortium, we implemented the OMOP CDM on the UK Electronic Healthcare Record database—The Health Improvement Network (THIN).

Objective The aim of the study was to evaluate the implementation of the THIN database in the OMOP CDM and explore its use for active drug safety surveillance.

Methods Following the OMOP CDM specification, the raw THIN database was mapped into a CDM THIN database. Ten Drugs of Interest (DOI) and nine Health Outcomes of Interest (HOI), defined and focused by the OMOP, were created using the CDM THIN database. Quantitative comparison of raw THIN to CDM THIN was performed by execution and analysis of OMOP standardized reports and additional analyses. The practical value of CDM THIN for drug safety and pharmacoepidemiological research was assessed by implementing three analysis methods: Proportional Reporting Ratio (PRR), Univariate Self-Case Control Series (USCCS) and High-Dimensional Propensity Score (HDPS). A published study using raw THIN data was selected to examine the external validity of CDM THIN.

Results Overall demographic characteristics were the same in both databases. Mapping medical and drug codes into the OMOP terminology dictionary was incomplete: 25 % medical codes and 55 % drug codes in raw THIN were not listed in the OMOP terminology dictionary, representing 6 % condition occurrence counts, 4 % procedure occurrence counts and 7 % drug exposure counts in raw THIN. Seven DOIs had <0.3 % and three DOIs had 1 % of unmapped drug exposure counts; each HOI had at least one definition with no or minimal (≤0.2 %) issues with unmapped condition occurrence counts, except for the upper gastrointestinal (UGI) ulcer hospitalization cohort. The application of PRR, USCCS and HDPS found, respectively, a sensitivity of 67, 78 and 50 %, and a specificity of 68, 59 and 76 %, suggesting that safety issues defined as known by the OMOP could be identified in CDM THIN, with imperfect performance. Similar PRR

Electronic supplementary material The online version of this article (doi:10.1007/s40264-012-0009-3) contains supplementary material, which is available to authorized users.

X. Zhou (✉) · S. Murugesan · Q. Liu · B. Cai · A. Bate
Epidemiology, Worldwide Safety Strategy, Pfizer,
219 E 42nd Street, Mail Stop 219/9/01,
New York, NY 10017, USA
e-mail: xiaofeng.zhou@pfizer.com

H. Bhullar
Cegedim Strategic Data Medical Research Ltd, London, UK

C. Wentworth
Analytic Consulting Solutions, Wakefield, RI, USA

A. Bate
School of Information Systems, Computing and Mathematics,
Brunel University, London, UK

A. Bate
Division of Clinical Pharmacology,
NYU School of Medicine, New York, NY, USA

scores were produced using both CDM THIN and raw THIN, while the execution time was twice as fast on CDM THIN. There was close replication of demographic distribution, death rate and prescription pattern and trend in the published study population and the cohort of CDM THIN.

Conclusions This research demonstrated that information loss due to incomplete mapping of medical and drug codes as well as data structure in the current CDM THIN limits its use for all possible epidemiological evaluation studies. Current HOIs and DOIs predefined by the OMOP were constructed with minimal loss of information and can be used for active surveillance methodological research. The OMOP CDM THIN can be a valuable tool for multiple aspects of pharmacoepidemiological research when the unique features of UK Electronic Health Records are incorporated in the OMOP library.

1 Background

The use of longitudinal observational data is a critical component of drug safety surveillance. The observational databases, both Electronic Health Records (EHRs) and transactional claims, are now widely and routinely used in pharmacoepidemiology [1]. With increased capability to access and link data and the inherent limitations of any single database [2], there has been an increased interest in the potential of multiple observational database analyses to understand the safety profile of marketed drugs. Studies using multiple databases might be undertaken to augment sample size, increase population heterogeneity, cross-validate results, or for other reasons. Observational databases store different information and have diverse data structures (e.g. different fields, field names, coding conventions, and terminologies for classifying drugs and outcomes) [3]. Studies using multiple databases in addition to methodological challenges, also therefore provide practical challenges in performing analyses across heterogeneous datasets. With this focus on the studies using multiple databases, there is renewed need and potential for considering alternative data access models to optimize analysis capability while ensuring patient privacy. The use of decentralized or 'distributed' models, where data is held remotely and central access of data is solely limited to summarized data, is intuitively an appealing approach. One such approach to introduce standardization, and therefore increased efficiencies, into a decentralized model is to convert all databases held remotely into a standard format, a 'Common Data Model' (CDM). When data are stored in such a CDM, a single query can be executed in parallel with minimal intervention across multiple databases, all with different underlying structure and properties.

The Observational Medical Outcome Partnership (OMOP) is a public private partnership, initiated to conduct research activities to assess the contribution and utility of using multiple observational data sources to identify and evaluate safety issues associated with marketed drugs [4, 5]. It has developed and tested 13 methods to apply to a network of central and distributed data sources for drug safety and effectiveness questions [4, 5]. One important feature of the OMOP research model is the OMOP CDM, a common structure and framework for organizing and standardizing observational databases. Such a model allows a range of standard queries and analytic methods developed centrally at the OMOP to be run seamlessly in both the distributed network and a centralized environment, potentially allowing the provision of fast quality-assured results. Other CDM structures have been proposed for observational databases for use in active surveillance activities, including the CDMs used by the EU-ADR [6] and the US FDA's Mini-Sentinel [7].

Effective conversion of a source database into a CDM underpins the value of both a CDM-based centralized environment of databases and also a distributed network of databases for standardized analyses. The published OMOP experience with the CDM has been confined to US databases, and primarily to claims transaction databases. Consisting of anonymous demographic, medical and prescription information at patient level, The Health Improvement Network (THIN) database provides longitudinal EHRs directly collected from general practitioners (GPs) in routine UK healthcare systems. THIN is an extensively validated database [8] that has been used very widely for pharmacoepidemiological research [9–11]. We chose to map the THIN database into the CDM THIN database, following the CDM specification provided by OMOP.

Some previous work has looked at challenges and opportunities associated with implementing and using a CDM, including some work on the OMOP CDM. Converting ten US observational databases into the OMOP CDM, Overhage et al. [12] assessed the feasibility, fidelity and resources required for implementing OMOP CDM across these databases and related standard terminologies to support method development for active drug surveillance. Two OMOP tools were used to validate the OMOP CDM quality. Eleven analytic methods were executed across all OMOP CDM instances to measure performance. Assessment of the OMOP CDM quality focused, at aggregated level, on the completeness of the concept mapped for conditions and medications, and the fidelity issues identified by the OMOP tools across the databases. The validation of CDM use for drug active surveillance was limited to the technical performance of the methods on the CDM to determine the execution time. No external

validation of results were performed. Reisinger et al. [13] developed a CDM and mapped two US observational databases into that CDM. Detailed descriptive statistics were used to validate the mapping quality and data aggregation process. Analysis focused solely on one outcome cohort (acute myocardial infarction [AMI]) and one drug (rofecoxib) cohort, and the execution of a standard-analysis program against the transformed data in each database to assess the performance of utilizing the CDM for drug safety analysis. While the authors concluded that the CDM was valid, the generalizability of this result is limited as, in addition to restricting analysis to a sole drug-outcome pair, neither active surveillance methods nor OMOP tools were used in the CDM evaluation process. Van Le et al. [14] then evaluated the performance of a semi-automated approach for risk estimates using the same CDM and US observational databases as used in Reisinger et al. [13]. They compared the risk estimates generated by this approach with those based on the published epidemiological evaluation studies using various observational databases, and demonstrated the compatible results. With a focus on the method development, their study assumed the validity of the CDM.

The objective of this study is to assess the appropriateness of the OMOP CDM structure for a UK EHR database (THIN) by both internal and external validation and its potential use for safety surveillance by the implementation of three statistical methods. We aim to give insights to the value and challenges associated with safety surveillance applied to a standardized database structure held centrally or accessed remotely.

2 Methods

2.1 Implementing The Health Improvement Network (THIN) Database onto the Observational Medical Outcome Partnership (OMOP) Common Data Model (CDM)

2.1.1 The OMOP CDM

The OMOP CDM consists of nine data tables (Fig. 1). There were seven core data tables ('Person', 'Condition Occurrence', 'Drug Exposure', 'Procedure Occurrence', 'Visit', 'Observation' and 'Observation Period'), and two data tables ('Condition ERA' and 'Drug ERA') aggregated, respectively, from condition occurrence and drug exposure data. A key feature of the OMOP CDM is to use a standard 'concept', which the OMOP describes as "the basic unit of information" and is a unique identifier in the OMOP dictionary [15]. Concepts are stored in the OMOP terminology dictionary so that any conceptual construct in a CDM data table is related (or mapped) to the appropriate corresponding standard concept in the dictionary (Fig. 1). This allows the integration and standardization of terminologies across multiple data sources for OMOP standardized analysis.

2.1.2 Mapping THIN into CDM THIN

THIN data, provided by Cegedim Strategic Data Medical Research (CSD MR), London, UK, are organized into four main categories: demographic, diagnoses, prescribing and additional health information [16]. The data in each of the four data tables were transferred into the OMOP CDM THIN data tables. The relationship between raw THIN and OMOP CDM THIN are presented in Fig. S1 (Online Resource 1). Additional details of the database conversion and evaluation are provided in the Electronic Supplementary Material (Online Resource 1). The THIN version with last collection date on 29 July 2009 was used for this analysis.

Following the OMOP CDM specification (version 2) [15], the following procedures were used to map the raw THIN database into the OMOP CDM THIN.

1. *Transforming the raw THIN contents to conform to the data structures of the OMOP CDM tables:* This includes direct mapping a source code into CDM THIN with the same name or a new name with a semantically identical interpretation; converting or deriving a source code before mapping when the data element could not be directly transposed. If conversion could not be conformed, data in specific fields were not transferred. Mapping approaches used for each variable in CDM THIN were described in Fig. S2 (Online Resource 1).
2. *Augmenting the source data with corresponding standard concept codes in the OMOP terminology dictionary:* This allows source codes such as medical codes (Read codes) and sex to look up the matched standard concept IDs in the dictionary using the OMOP source-to-concept mapping system in the OMOP vocabulary library.
3. *Aggregating drug and condition eras within a predefined persistence window, which has been described in detail by the OMOP* [15, 17, 18]. For this study, drug usage was aggregated into one drug era if the start of the subsequent prescription occurred within 30 days (persistence window) of the end of the previous prescription. Similarly, the condition was aggregated into one condition era if the start of the subsequent condition occurred within 30 days (persistence window) of the end of the previous condition occurrence. The selection of the persistence window was based on the options recommended by the OMOP [15].

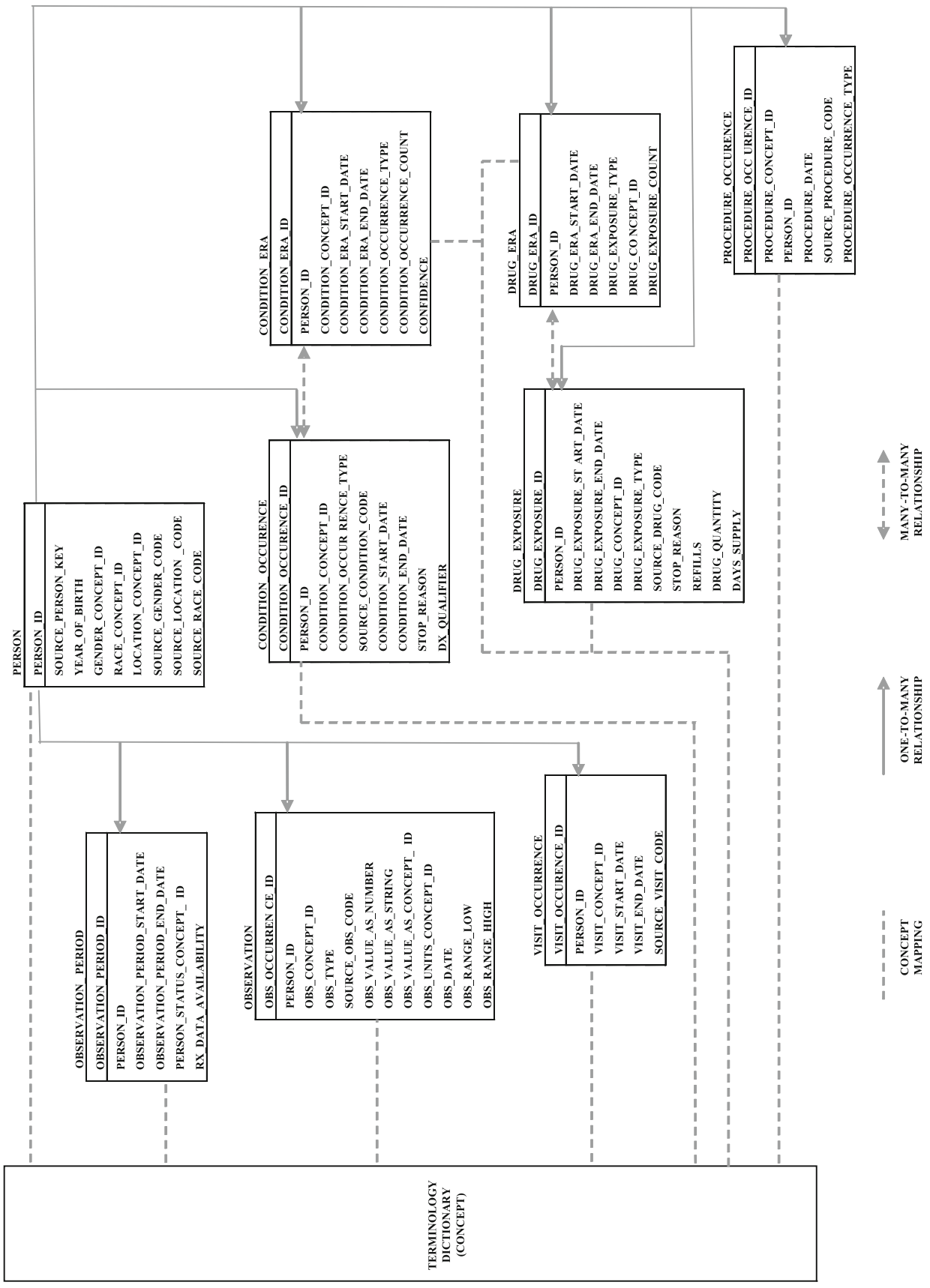

Fig. 1 Database model of OMOP (Observational Medical Outcome Partnership) Common Data Model. *Dx* diagnosis, *OBS* observation, *Rx* prescription

4. *Documentation*: procedures, assumptions and criteria used for guiding the data Extract, Transform, and Load (ETL) during the mapping process were documented using the standard ETL specification provided by the OMOP.

2.1.3 *Creating Health Outcomes of Interest (HOI) and Drugs of Interest (DOI) Cohorts*

The OMOP defined and focused analyses on nine Health Outcomes of Interest (HOI) and ten Drugs of Interest (DOI) [Table 1], each HOI with multiple definitions of varying specificity. We used the standard SAS Software (SAS Institute Inc., Cary, NC, USA) programs provided by the OMOP to run on CDM THIN and to generate the HOI and DOI cohorts.

2.2 Internal Validity Evaluation

2.2.1 *Assessing the Quality of CDM THIN*

The OMOP developed tools with the intent of standardized report and quality assessment across different observational databases in the OMOP CDM. These standardized tools as well as the independent programming were employed to evaluate the data mapping process between raw THIN and CDM THIN. The goal was to ensure that the data transformation did not introduce errors or inconsistencies in the underlying data. The OMOP tools used for this purpose were (i) The Observational Source Characteristics Analysis Report (OSCAR) [19, 20], a tool to provide a summary of data characteristics for the key fields in the nine CDM THIN data tables; this tool facilitates detection of data quality issues in the overall population; (ii) The Natural History Analysis (NATHAN) report [19, 21], a data characterization tool for a specific subpopulation, intended to reveal unusual patterns within these subpopulations, in particular, focusing on the HOIs and DOIs predefined by OMOP; and (iii) Generalized Review of OSCAR Unified Checking (GROUCH) [19, 22], a data anomaly detection tool, which was used to report if there were incomplete, implausible and suspicious data (e.g. year of birth before 1900, drug supply days <0, etc.) across all tables in CDM THIN. Additional independent programming using raw THIN and CDM THIN was performed to compare with the findings from the OMOP standard reports, and to identify and further understand any data issues beyond the coverage of these descriptive reports.

Read codes and Multilex IDs are medical classification and drug terminology codes, respectively, used in the THIN database. The OMOP had mapped Read codes to SNOMED (a multiaxial, hierarchical medical classification system), and Multilex IDs to RxNorm (a system providing normalized names and unique identifiers for medicines and drugs). SNOMED and RxNorm codes had been assigned concept IDs by the OMOP in the OMOP terminology dictionary [23]. Read codes in raw THIN contain both condition and procedure codes that are mapped into condition and procedure tables, respectively. The details of the OMOP terminology mapping have been described elsewhere [12, 15, 19]. The completeness of the mapping of the Read codes and Multilex IDs into the OMOP terminology library was specifically assessed by analysis of descriptive statistics and extensive consultation with a source data expert as well as the OMOP. As one code could be associated with multiple counts of condition or procedure occurrence, or drug exposure per patient, we also examined those unmapped codes associated with the most counts of condition and procedure occurrences, as well as drug exposure. This allowed us to further investigate the characteristics of the unmapped codes. Special attention was given to investigating the completeness of the codes used for creating the nine HOI cohorts and ten DOI cohorts through detailed qualitative and quantitative evaluation as

Table 1 Health Outcomes of Interest and Drugs of Interest cohorts [39]

	Health Outcomes of Interest	Drugs of Interest
1	Angioedema	ACE inhibitor
2	Aplastic anemia	Amphotericin B
3	Acute liver failure	Antibiotics: erythromycins, sulfonamides and tetracyclines
4	Acute renal failure	Antiepileptics: carbamazepine, phenytoin
5	Bleeding	Benzodiazepines
6	Hip fracture	β-blockers
7	Acute myocardial infarction	Bisphosphonates
8	Mortality after myocardial infarction	Tricyclic antidepressants
9	Upper gastrointestinal ulcer hospitalization	Typical antipsychotics
10	Hospitalization[a]	Warfarin

OMOP Observational Medical Outcome Partnership

[a] Hospitalization cohort was not used in the focused analysis by the OMOP and this study (i.e. for applying the active surveillance methods)

they were the focus of the method testing in this study. A stepwise algorithm was used to identify unmapped Read codes and Multilex IDs for each HOI or DOI. This was done by (i) identification of the discordant codes not mapped in the OMOP terminology dictionary: for outcomes, identifying all Read codes in the OMOP definitions for each HOI, going up two levels in the hierarchy and then searching all Read codes in raw THIN linked to these high-level OMOP Read codes; for drugs, performing text mapping of drug names in the OMOP definitions for each DOI into the raw THIN; (ii) manually reviewing the results of the discordant codes, confirming with a source data expert and consulting with an internal physician.

2.2.2 *Assessing the Value of the OMOP CDM THIN in Executing Analyses*

To evaluate the value of our CDM THIN implementation, we tested and evaluated the OMOP's standardized analytic methods. The Proportional Reporting Ratio (PRR) method was one method selected from the OMOP method library for tractability testing. It is a simple disproportionality (DP) analysis method that is a standard quantitative approach for analysing data from Spontaneous Reporting Systems (SRSs) for drug safety signal detection [24, 25]. There have been some efforts made to implement the PRR method to longitudinal healthcare records, and the nature of these records means the underlying contingency table can be defined and populated using different approaches.

Consider a 2 × 2 contingency table:

	Outcome of interest	Other outcomes
Drug of interest	a	b
Comparator drugs	c	d

$$\mathrm{PRR} = [a/(a+b)]/[c/(c+d)]$$

where, in this study, a = the number of incident target conditions that occur during target drug eras; b = the number of incident non-target conditions that occur during target drug eras; c = the number of incident target conditions that occur during non-target drug eras; and d = the number of incident non-target conditions that occur during non-target drug eras.

In brief, the PRR method was applied to CDM THIN using the standard OMOP program with the incident condition type (first occurrence) and SRS approach [24]. Crude estimates without adjustment for potential confounders were produced. In addition to the PRR score, the PRR 95 % LBCI > 1 (the lower bound of the two-sided 95 % confidence interval being greater than 1) was used as a decision criterion for defining positive test results [26].

We assessed concordance of the PRR results with the OMOP 'Ground Truth', the reference set of 53 drug-outcome pairs selected by the OMOP for which extensive literature searching had been performed for assessing true drug-event association [26]. Among the 53 pairs, 9 were considered as known safety associations and 44 as negative controls. We also compared the PRR results on CDM THIN with those on raw THIN and made the comparison to the OMOP empirical testing results across ten US databases [26].

In addition, we also tested two other methods that have been extensively investigated for traditional pharmacoepidemiological studies—the Univariate Self-Case Control Series (USCCS) [27, 28] and High-Dimensional Propensity Score (HDPS) [29, 30]—to further assess the readiness and utility of CDM THIN to handle the standardized and validated methods provided by the OMOP for safety surveillance. The results of these methods were also compared with the OMOP 'Ground Truth' and the OMOP empirical testing results across ten US databases.

The self-control case series method automatically controls for time-fixed covariates by only including cases in analyses, which are then used as their own controls [31]. The USCCS model considers one drug and one outcome combination, which is assumed to arise from a non-homogeneous Poisson process. We utilized the SAS codes and parameter setting (Appendix S1; Online Resource 1) of the USCCS method provided by the OMOP and calculated the relative risk (RR) scores for incident cases [32]. HDPS is a multistep algorithm to empirically identify candidate covariates, prioritize covariates, and integrate them into a propensity-score-based confounder adjustment model for minimizing the confounding in observation data. The details involved in this approach have been described elsewhere [29, 30]. This study calculated the propensity scores and adjusted odds ratios (OR) by using the standard OMOP HDPS program and parameter setting (Appendix S1, Online Resource 1) [32].

All three methods were tested against the ten single DOI definitions. The OMOP had defined HOIs with multiple definitions. A single specific definition (Appendix S2; Online Resource 1) for each of the nine HOIs was selected for this method testing based on minimal information loss during the mapping process of the Read codes used for creating each cohort.

2.3 External Validity Evaluation

We searched all 241 publications listed on the CSD MR website [33] for any full publications that analysed any of the ten DOI and nine HOI combinations, and focused on

pairs for analysis by the OMOP and thus for us in this study. These criteria resulted in a single publication by Hardoon et al. in 2011 [34]. Additional analyses were performed to evaluate if it was possible to replicate the results of Hardoon et al. using CDM THIN.

Specifically, as far as possible, the same inclusion and exclusion criteria used in the publication were applied to create a cohort of AMI patients in CDM THIN between 1991 and 2002, aged 35 years or over at the time of AMI diagnosis, and survival of at least 3 months after the AMI. The proportion of patients who died from 3 months to 3 years after the first AMI diagnoses was calculated; and the proportion of patients who were prescribed the four drugs (lipid-regulating drugs, β-blockers, ACE inhibitors and antiplatelet medications) within 3 months of AMI diagnoses during 1991–2002 was calculated. These results were compared with those in the study by Hardoon et al. [34].

The implementation of CDM THIN mapping, as well as all analyses for internal and external validity evaluation, was done using SAS/SQL of SAS version 9.2. Both raw THIN and OMOP CDM THIN datasets are stored internally under secured UNIX server as SAS datasets.

3 Results

3.1 Mapping Raw THIN into CDM THIN

Nine OMOP CDM THIN data tables were generated, including seven core data tables and two data tables, from aggregating the condition occurrence and drug exposure data. A total of 7,109,761 patients with acceptable flags in raw THIN were transferred into CDM THIN with an average age of 45 years. Nine HOIs with multiple definitions and ten DOI cohorts in CDM THIN were created using the definitions of HOI and DOI and the standard SAS programs provided by OMOP. It took approximately two full-time person-years to complete this database conversion and validation.

3.2 Internal Validity

3.2.1 Overall Mapping Quality

All CDM-required data elements that are available in raw THIN were completely mapped into Person, Visit and Observation Period data tables. Four CDM data tables (Condition Occurrence, Drug Exposure, Procedure Occurrence and Observation), however, had incomplete mapping, which is described in Sects. 3.2.2 and 3.2.3. Overall demographic characteristics were identical across the two databases, including a 100 % match on age, sex, patient status and regional distribution, despite different mapping approaches for these four variables. Figure 2, a database model heat map, provides an overview of the field-by-field mapping quality. Selected findings from OSCAR and GROUCH are included in Tables S1 and S2 (Online Resource 1).

3.2.2 Overall Mapping Quality of Medical and Drug Codes

There was incomplete mapping of medical and drug codes into the OMOP standard terminology dictionary (Table 2): 25 % Read codes and 55 % Multilex IDs in raw THIN were not listed in the OMOP terminology dictionary, representing 6 % of condition occurrence counts, 4 % of procedure occurrence counts and 7 % of drug exposure counts in raw THIN.

Grouping the unmapped condition codes by condition class on the first (i.e. the highest hierarchical) level of Read code, we found that the classes with the most commonly unmapped condition codes ($n > 300$) included 'External morbidity and mortality', 'Neoplasm', 'Injury and poisoning' and 'Mental disorders'. Among all condition occurrences associated with the unmapped condition codes, 'Symptoms, signs and ill-defined conditions', 'Circulatory system diseases', and 'Skin and subcutaneous tissue diseases' accounted for the largest proportion of each condition class (>15 % per class). More extensive details of the findings can be found in Fig. S3 (Online Resource 1). We also checked for the most common unmapped condition codes based on the lowest level of terminology of Read code (7th level). Table 3 lists the 20 most common unmapped condition codes that represent 59 % of all unmapped condition occurrence counts. The results indicate that some of the unmapped codes might be associated with some common diagnosis, such as hypertensive disease and ischaemic heart disease, etc. Further inspection shows that these unmapped codes seem unrelated to the HOI definitions, with the exception of angina pectoris, which can be a code used in an AMI definition but is not included in the OMOP definition of AMI used in this study. We took the similar approaches to evaluate the mapping quality of procedure codes. The results revealed that the majority of the procedure occurrence associated with unmapped codes were not involved in a specific diagnosis or therapeutic procedure. Overall, 74 % of all unmapped procedure occurrence counts were associated with four specific procedure codes: 'Had a chat to patient', 'Medication requested', 'Discharged from hospital' and 'Depression screening using questions'. There was a loss of granularity during Read code terminology mapping: Read code terminology in raw THIN has seven levels; however, the OMOP dictionary only allowed for the first five digits of

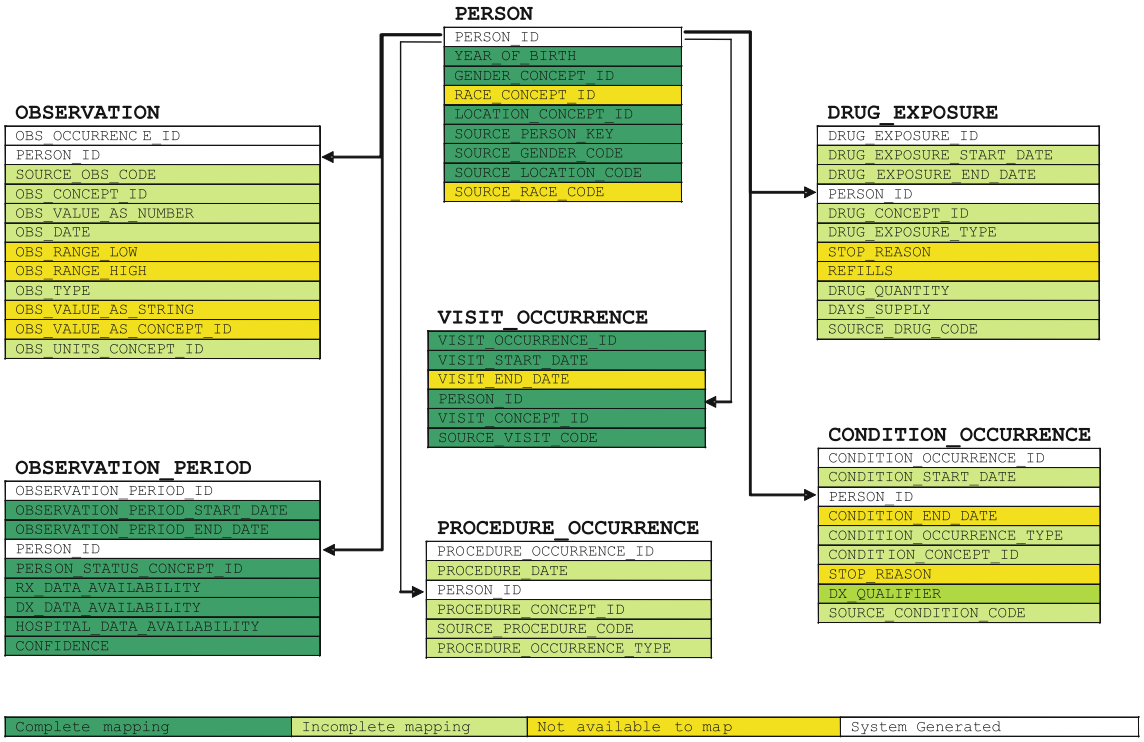

Fig. 2 Database heat map: overall mapping quality of the THIN (The Health Improvement Network) database in the OMOP (Observational Medical Outcome Partnership) Common Data Model. *Dx* diagnosis, *OBS* observation, *Rx* prescription

the Read code to be mapped. Consequently, a Read code listed at six- or seven-digit level, needed to be mapped to a higher level term—fifth level (e.g. G30..11 is mapped to G30..).

Grouping the unmapped drug codes into each drug class, on the first (i.e. the highest hierarchical) level of British National Formulary codes that are linked with each Multilex ID, we found that most commonly unmapped drug classes ($n > 2000$) included 'Skin', 'GI system', 'Obstetrics and gynaecology and urinary tract disorders' and 'Nutrition and blood'. Among all drug exposures associated with the unmapped drug codes, 'Skin', 'Endocrine system', and 'Nutrition and blood' accounted for the largest proportions of each drug class (>10 % per class). More extensive details of the findings can be found in Fig. S4 (Online Resource 1). We examined the unmapped drug codes for the ten DOIs and found that the majority of the drugs were US brands. We noted that a single brand drug (e.g. warfarin) could be both mapped and unmapped in the OMOP terminology dictionary depending on its dose strength. Table 3 lists the 20 most common unmapped drug codes that made up 24 % of all unmapped drug exposure counts. On inspection it seems that these unmapped codes are unrelated to the DOI definitions and the majority are not traditional medications. However, the number of vaccines in this list makes it clear that further mapping work is needed for the OMOP CDM to be of value to UK EHR databases for vaccine vigilance and surveillance activities by ensuring mapping of vaccine Multilex IDs into the terminology library.

3.2.3 *Mapping Quality of Medical and Drug Codes for HOIs and DOIs*

In order to assess the mapping of diagnosis and drug codes for HOIs and DOIs, the proportion of unmapped condition occurrence counts or drug exposure counts within each HOI or DOI, respectively, was calculated. Figure 3 shows that among ten DOIs, seven had <0.3 % and three had 1 % of drug exposure counts unmapped in CDM THIN. Among nine HOIs with multiple definitions per HOI, each HOI had at least one definition with no or minimal issues (≤0.2 %) with unmapped condition occurrence counts, except for the upper gastrointestinal ulcer hospitalization cohort. Figure 3 shows the proportion of unmapped condition occurrence counts for the definitions of each HOI used in this study. HOI cohorts involving procedure codes in this study were not generated because (i) specific procedure codes in raw THIN were unmapped because of an absence of standard concept IDs in the terminology dictionary for these codes; or (ii) data for some specialized procedures in HOIs were not collected in raw

Table 2 Performance of mapping drug and medical terminologies in raw THIN into the Observational Medical Outcome Partnership standard terminology dictionary

Terminology type	No. of unmapped codes[a]	Proportion of the unmapped codes in raw THIN (%)	Exposure counts or occurrence counts associated with unmapped codes[a]	Proportion of the unmapped counts in raw THIN (%)
Multilex ID (drug codes)[b]				
Drug	18,776	50.34	39,205,281	6.6
Homoeopathic/herbal remedy	430	1.15	47,957	0.01
Appliances/others	306	0.82	30,761	0.01
No information available for product set	901	2.42	576,945	0.10
Total	*20,413*	*54.73*	*39,860,945*	*6.71*
Read codes (medical codes)[c]				
Condition	4,169	6.98	8,665,598	2.41
Unspecified condition	5,897	9.87	11,317,861	3.15
Procedure	2,676	4.48	13,849,888	3.85
Other	2,168	3.63	36,228,558	10.08
Unknown	31	0.05	443	0.00
Total	*14,941*	*25*	*70,062,348*	*19.50*

Source: THIN Data Guide For Researchers, Cegedim Strategic Data Medical Research, UK, July, 2009

BNF British National Formulary, *THIN* The Health Improvement Network

[a] Unique code/count with acceptable flag in raw THIN

[b] The BNF's Multilex ID is a hierarchical terminology with pre-defined groups of medicinal products. All medicinal drugs (made up of 15 drug classes) were grouped together, excluding herbals and homeopathic, which were classified separately. All other products were grouped together in 'Appliances/others' except those classified by the BNF as having no information available, which were also kept as a separate category. The 'Appliances/others' category includes cough and cold preparations, chiropody products, urinary catheters and trusses, etc

[c] The Read codes were divided into five major categories using the highest hierarchical level defined in the Read code. 'Condition' was made up of 19 condition classes representing all classes of conditions, except those explicitly classified in the hierarchy as 'unspecified', which were considered separately. All 'Procedure' codes were grouped together, except those that were classified explicitly as occupations, historic symptoms or administration codes; these codes were kept as a single group. As the fifth group, codes that were unknown and considered as non-standard by the terminology were kept separate

THIN as primary care GPs may not record secondary care procedures in THIN routinely (e.g. bone marrow aspiration or biopsy).

3.2.4 Mapping of Laboratory and Other Additional Health Data

The OMOP used Logical Observation Identifiers Names and Codes (LOINC) as standard terminology for mapping laboratory data. In raw THIN, laboratory data were not coded in LOINC, but rather Additional Health Data (AHD) codes. Furthermore, many of the laboratory units in raw THIN were not reflected in the OMOP terminology dictionary. We only mapped the ten laboratory tests that were specifically required in the HOI definitions used in this study into CDM THIN. Physical examination data such as body mass index, height, weight, etc., and lifestyle data such as smoking, alcohol, etc., in the AHD table in raw THIN were not transferred into CDM THIN due to a difference in data structure and the lack of available concept IDs in the OMOP terminology dictionary.

3.2.5 Aggregated Drug Exposures and Condition Occurrences

The aggregation process for creating eras significantly reduced the overall number of drug exposure counts per person, but not condition occurrence counts. In the Drug_ERA table, the mean number of drug eras per person is 32, indicating a 66 % reduction of the mean number of drug exposures per person compared with the mean number of drug exposures per person ($n = 93$) in the Drug_Exposure table. In the Condition_ERA table, the mean number of condition eras per person is 23, indicating a 12 % reduction of the mean number of condition occurrences per person compared with the mean number of condition occurrences per person ($n = 26$) in the Condition_Occurrence table.

Table 3 Top 20 unmapped condition occurrence counts and drug exposure counts

Condition description	Condition occurrences per unmapped code among total unmapped condition occurrences (%)[a]	No. of condition occurrences per unmapped condition code	Generic name of product	Drug exposures per unmapped code among total unmapped drug exposures (%)[b]	No. of drug exposures per unmapped drug code
Hypertensive disease	7.55	653,879	Aqueous CRM 4 kilo(s)	3.70	1,450,990
[D]Rash and other non-specific skin eruption NOS	7.16	620,585	Emollient + hypoallergenic lanolin and hydroxybenzoates CRM	3.32	1,299,943
Thrush	4.58	397,229	Influenza inactivated surface antigen vaccine	1.34	526,635
Angina pectoris	4.50	389,712	Influenza inactivated split virion vaccine ten 0.5 mL pre-filled syringe	1.15	451,920
[D]Insomnia NOS	3.59	311,186	Blood glucose testing strip product	1.15	449,792
Gastroenteritis	3.47	301,027	Vitamin B compound strong tablets, 28 tablets	1.13	442,748
Ischaemic heart disease	3.02	261,815	Glucose enzymatic test strip product	1.12	440,388
Urticaria	2.91	252,097	Sterile spec 10 dressing pack, 12 pack	1.08	423,787
[D]Vertigo NOS	2.55	220,681	Permeable non-woven BP 88 synth ADH tape 2.5 cm	1.08	422,412
Leg ulcer NOS	2.46	213,414	Blood glucose testing strip product	1.06	414,019
Sebaceous cyst – wen	2.20	190,766	Soya bath oil	1.01	395,438
Chronic obstructive pulmonary disease	2.07	179,601	Rabeprazole GR tablets 10 mg, 7 tablets	0.99	388,135
[D]Fever NOS	2.03	176,268	Blood glucose testing strip product	0.83	325,107
Candidiasis	2.00	173,498	Vitamin BPC capsules, 250 capsules	0.81	317,278
Varicose veins of the legs	1.79	154,797	Spacer/holding chamber device type 2	0.78	304,862
Overdose of drug	1.66	143,873	Blood glucose testing strip product	0.76	299,243
Influenza	1.53	132,712	Influenza inactivated split virion vaccine	0.68	267,999
Folliculitis	1.41	121,893	Blood glucose testing strip product	0.68	267,088
Chronic obstructive airways disease	1.39	120,553	Oral rehydration salts oral powder	0.68	266,254
Enterobiasis – threadworm	1.20	103,633	Simple LINC 200 mL	0.65	256,659
Total	*59.08*			*24.00*	

[D] working diagnosis, *NOS* not otherwise specified

[a] 'Total unmapped condition occurrences' are the total number of condition occurrence counts associated with all unmapped condition codes

[b] 'Total unmapped drug exposures' are the total number of drug exposure counts associated with all unmapped drug codes

3.2.6 Proportional Reporting Ratio Method

PRR scores and PRR 95 % LBCI scores were produced by the PRR method testing on the 53 drug-outcome pairs in the OMOP 'Ground Truth' set. Seven out of nine positive pairs with known safety issues had a PRR >1 (Table 4), indicating that there were more observed occurrences of the combination of interest than expected under a null hypothesis of no association between drug and HOI. Twenty-seven pairs out of 44 negative controls had a PRR <1, indicating that these pairs had no association between a DOI and HOI. In addition, six out of nine known positive sets had a PRR 95 % LBCI >1, and 30 out of 44 negative controls had a PRR 95 % LBCI <1.

We compared our results with the OMOP's analysis of DP methods based on pooled scores across ten US databases (Table 5) [26]. The OMOP presented an overall DP score, not specific to the PRR, but the results could be anticipated to be comparable [35, 36], particularly as standard DP methods do not account for the longitudinal nature of EHRs [37]. The PRR on CDM THIN offers better sensitivity (67 vs. 44 %) and lower specificity (68 vs. 73 %) compared with the OMOP DP testing results.

Using the same definitions of HOI and DOI for testing PRR on CDM THIN, we implemented the PRR method on raw THIN. Table 4 shows that the results on both CDM THIN and raw THIN are identical except that the PRR on raw THIN picked up two more true negative pairs when applying the criteria of PRR 95 % LBCI <1. The execution time of the PRR on CDM THIN was much shorter compared with raw THIN, i.e. 5.4 versus 11.7 h.

Table 4 Comparison of Proportional Reporting Ratio method results with the OMOP Ground Truth: CDM THIN vs. Raw THIN

Threshold	N (pairs)	
	CDM THIN	Raw THIN
Out of nine known positive sets in the OMOP ground truth set		
PRR >1	7	7
PRR 95 % LBCI >1	6	6
Out of 44 negative controls in the OMOP ground truth set		
PRR <1	27	27
PRR 95 % LBCI <1	30	32

CDM Common Data Model, *LBCI* lower bound confidence interval, *OMOP* Observational Medical Outcome Partnership, *PRR* proportional reporting ratio, *THIN* The Health Improvement Network

3.2.7 Univariate Self-Case Control Series Method and High-Dimensional Propensity Score Method

USCCS scores indicated that all nine pairs considered true positive sets by the OMOP had RR >1, and seven pairs demonstrated statistical significance (95 % LBCI >1). Twenty-six true negative pairs out of 44 OMOP negative controls were identified. Comparing the results of the OMOP USCCS method using ten US databases (Table 5), USCCS on CDM THIN offers much better sensitivity (78 vs. 44 %) but lower specificity (59 vs. 68 %), representing a false positive rate of 41 %. More than 70 % of false positive pairs, however, had RR <2.

Running the HDPS method on CDM THIN, we were able to get the results of 45 drug-outcome pairs that were compared with the 53 pairs of the OMOP 'Ground Truth'. The remaining eight pairs, including one positive pair and seven negative controls, in the OMOP reference set had no results generated due to 'insufficient data or no data' in some strata. Among the eight OMOP positive pairs, six had an adjusted OR >1, of which four pairs had 95 % LBCI >1. Out of 37 pairs of OMOP negative controls, 28 true negative pairs were identified. Comparing the results of the OMOP HDPS method using ten US databases (Table 5),

Table 5 Comparison of performance characteristics of three methods: CDM THIN vs. the OMOP CDM (ten databases)

Threshold	CDM THIN		OMOP CDM	
	Sensitivity	Specificity	Sensitivity	Specificity
PRR	0.67	0.68	0.44[a]	0.73[a]
PRR 95 % LBCI >1				
USCCS	0.78	0.59	0.44	0.68
RR >1 and LBCI >1 (a = 0.05)				
HDPS	0.50	0.76	0.56	0.82
OR >1 and LBCI >1 (a = 0.05)				

CDM Common Data Model, *DP* disproportionality, *HDPS* High-Dimensional Propensity Score, *LBCI* lower bound confidence interval, *OMOP* Observational Medical Outcome Partnership, *OR* odds ratio, *PRR* Proportional Reporting Ratio, *RR* relative risk, *THIN* The Health Improvement Network, *USCCS* Univariate Self-Case Control Series

[a] Results of DP method. PRR is one of the methods in the DP method class. PRR 95 % LBCI >1 means PRR >1 and LBCI >1 (a = 0.05)

HDPS on CDM THIN offers lower sensitivity (50 vs. 56 %) and specificity (76 vs. 82 %).

3.3 External Validity

The comparison between CDM THIN and the study by Hardoon et al. [34] revealed that our study had the same sex distribution (64 % male), with slightly younger age (mean age 66.7 vs. 67.9 years).

Comparing the proportion of patients who died from 3 months to 3 years after the first AMI diagnoses between the study by Hardoon et al. [34] and our study, we found that the death rate from CDM THIN analysis was slightly higher, but the trend in death rate over time (1991–2002) from both analyses was comparable (Appendix S3; Online Resource 1). Nevertheless, our study identified a larger number of patients who died as well as AMI patients, which could result from different practice selection in the newer version of the THIN database that we used as well as the differences in the algorithms to identify patients with AMI and patients who died between these two studies. Hardoon et al. [34] used the THIN database that covered a 15-year period from 1 January 1991 to 31 December 2005, and selected 218 practices that had at least 5 years worth of data; we used the THIN database from 1 January 1991 through 29 July 2009, with all 466 available practices.

For medication after AMI diagnosis, we observed that the prescription rates of all four drugs within 3 months of diagnosis of AMI were comparable, and the upward trend over time was almost identical between the two studies (Fig. 4).

Overall, we were able to replicate the conclusion that the death rate after the first AMI diagnosis decreased during the 12 years from 1991 to 2002, and this decrease was associated with the increased use of medication of lipid-regulating drugs, ACE-inhibitors, β-blockers and anti-platelet drugs.

4 Discussion

The internal validity evaluation of the THIN database in the OMOP CDM demonstrated the feasibility of data transformation with overall identical demographic characteristics across CDM THIN and raw THIN. Mapping medical, drug and laboratory codes into the OMOP standard terminology dictionary was incomplete; however, the information loss from the unmapped codes was minimal among the majority of HOI and DOI cohorts used in this study (Fig. 3). CDM THIN allowed application of standard definitions and assumptions defined by the OMOP for developing drug and condition cohorts as well as eras. Direct application of the OMOP standard analysis programs on THIN was then enabled. CDM THIN not only facilitates faster provision of analysis results, but also saves time for validating the results in comparison with the current approach using raw THIN, where the assumptions and criteria for defining a DOI/HOI, aggregating drug or condition episode, deriving variables, etc., are all embedded in individual analyses. The assessment of the technical performance of the PRR method on both CDM THIN and raw THIN indicated that the execution speed of the PRR on CDM THIN was twice as fast compared with raw THIN. This enhanced efficiency and timely assessment of safety issues using a CDM is in line with results from other studies [12, 13]. Our experiment added evidence that use of a CDM could make implementation of a distributed network of datasets with no necessity of centralized data

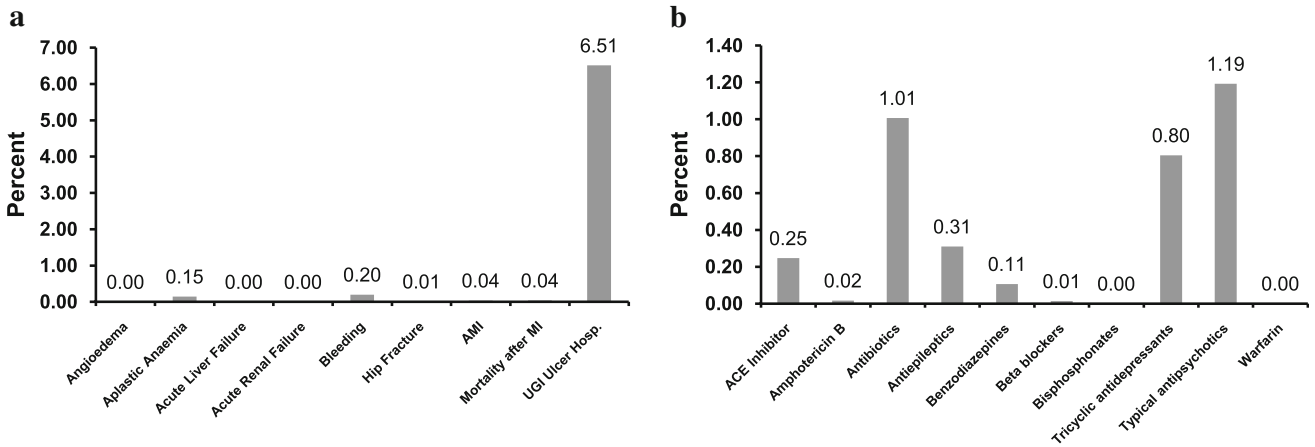

Fig. 3 Proportion of (**a**) unmapped condition occurrence counts within each Health Outcome of Interest and (**b**) drug exposure counts within each Drug of Interest. Note: The proportion is defined as the unmapped occurrence or exposure counts in this cohort divided by the sum of the mapped and unmapped occurrence or exposure counts in this cohort. *AMI* acute MI, *Hosp.* hospitalization, *MI* myocardial infarction, *UGI* upper gastrointestinal

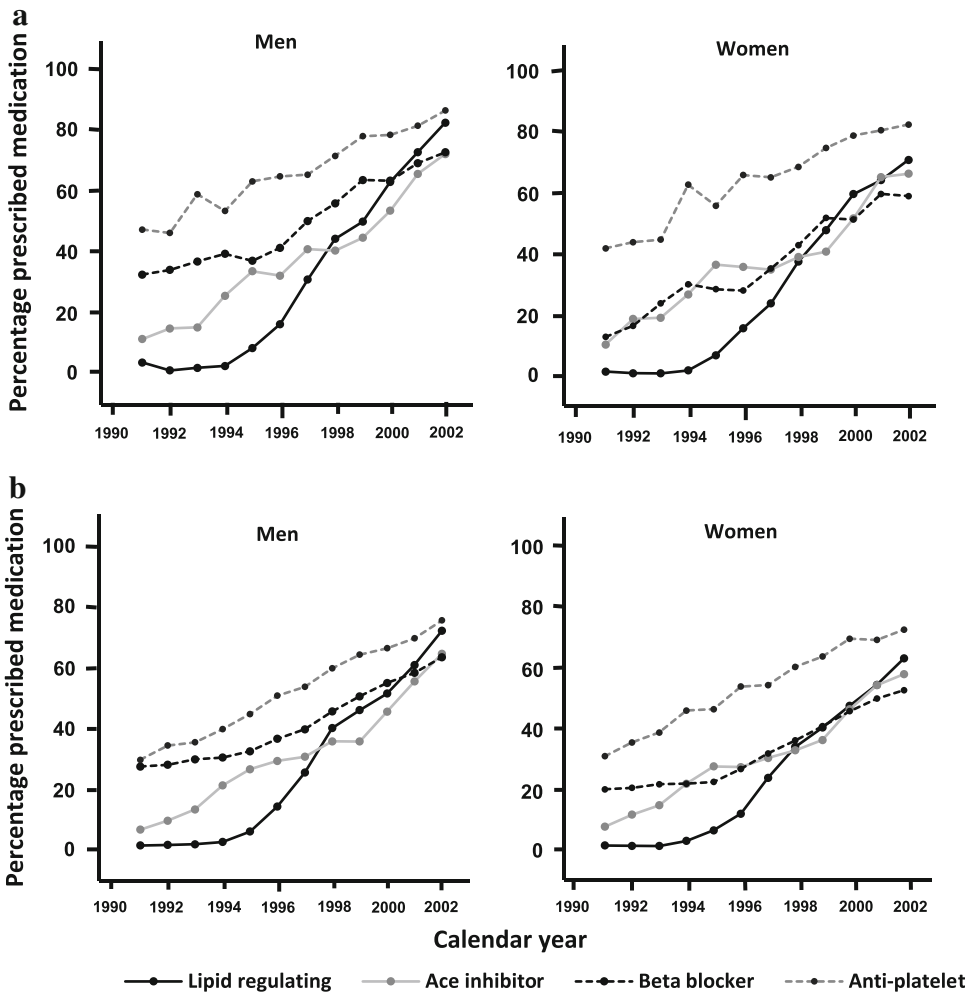

Fig. 4 Trend of prescriptions of the four drugs—published study (**a**) versus Common Data Model THIN (The Health Improvement Network) Evaluation Study (**b**)

collection a possibility. This is clearly of practical value in minimizing privacy concerns about the sharing of detailed patient level data. In the external validity evaluation, the results indicated a close replication of demographic distribution, death rate and prescription rates and trends for all four drugs within 3 months of diagnosis of AMI in the published study population and the cohort of CDM THIN, despite some differences between the analyses. While only one study, this nevertheless supports the position that the HOIs and DOIs focused in this study can be analysed and can provide results in line with those in the wider literature.

We strived to characterize Read codes and Multilex IDs that could not be mapped into the OMOP CDM with a special focus on those that may affect the creation of HOI and DOI cohorts. Our intent was not to assess the appropriateness of the OMOP definitions of HOI or DOI, but rather to understand if there are Read codes or Multilex IDs not included in the current OMOP definitions due to their unavailability in the OMOP dictionary. There is no established standard to measure if we captured 100 % unmapped codes using the algorithm in this study. However, a very close match of the PRR scores using the ten DOIs and nine HOIs on both CDM THIN and raw THIN, supports the mapping validity for creating these cohorts. Another important aspect of the internal validity is that we applied OMOP standard tools for the validation. It is worth noting that while these tools facilitated the validation process, OMOP did not provide instruction to determine if the identified warnings were legitimate anomalies and how the data issues should be corrected. User-specific evaluation of the findings to determine and interpret anomalies was required. We reviewed our findings with the data expert from CSD MR and decided on the appropriate action for correcting the data issues in CDM THIN. When a data inconsistency was judged to be due to mapping and converting data, we made the appropriate correction; when it was considered to be due to inherited properties of the original database, we kept the data so as to maintain maximum consistency with the source data.

Several key assumptions were used to map the data in raw THIN into CDM THIN, which could have impacted

the quality of mapping, and use of CDM THIN for analyses. For instance, to build an 'Observation_Period' that tracks the person's status during the study, we assumed (i) using the latter of start date using Vision software, Acceptable Mortality Reporting date and patient's registration date, as enrolment start date; (ii) when a patient transfers from one practice in THIN and enrols with another practice in THIN, this patient is seen as a new patient in the database; and (iii) using the last collection date as the enrolment end date; if a patient transferred out of the practice, the transfer-out date was used as the enrolment end date. The limitations of these assumptions include, firstly, that a patient may not be unique in this data table, and, secondly, that the last collection date of the database as enrolment end date will not be appropriate for a patient with missing transfer-out date. Duplicated patients are not common and missing transfer-out dates are rare occurrences in the data table, so the impact of these assumptions on analysis should be small. This is a general issue common to all studies conducted on UK EMRs. Two major assumptions were made to derive 'drug supply days'—prescription quantity divided by dose value—that was required for calculating 'end of the prescription': (i) we reviewed the dosage free text for the user decode dose value and obtained the appropriate numeric dose value if the text clearly indicated the daily dosage as 'once daily' or 'twice daily', etc. Otherwise, dose value was set to missing, thus 'days of supply' was missing; and (ii) for all missing 'days of supply', we imputed the most frequently occurred 'days of supply' value for that specific drug class. We note that the significant time would be required to update and validate the decode algorithm for every updated version of THIN and the approximation of missing days of supply may not always be appropriate for some specific drugs.

PRR scores were readily produced on CDM THIN. The PRR method produced comparable results when run on raw THIN (Table 4), and similar performance characteristics to the OMOP work on other data sources (Table 5). The tractability and ease of execution indicates the potential role in CDM screening for active surveillance. Comparison with the OMOP 'Ground Truth' suggested that the PRR approach on CDM THIN can highlight true associations, although at the expense of false positives. False negative results are also noted. Results could be produced for both HDPS and USCCS (Table 5), underscoring the value of a CDM, once the database conversion is completed, in allowing ready access to a suite of standardized and validated methods developed without a specific dataset in mind. While the performance characteristics of the methods varied in this analysis, we caution limited inference on these observed differences, other than an appreciation that no method will be perfect for surveillance. Future research on how to appropriately control confounding and determine an appropriate threshold for signal detection across the analysis methods is clearly needed. As for each of the three methods in this study, the results are based on a specific set of parameters and considered to be appropriate for the method in question [38]. Our intent was to test the utility of CDM for implementation of methods rather than to attempt to measure the method performance accurately. Additional testing of permutations of the parameter settings might be anticipated to improve the performance characteristics further, but given the limited size of the 'Ground Truth' test set we feel that any such evaluation would have limited value. Additionally, while the OMOP classifications of associations were based on extensive literature reviews they are well-studied examples of medications in widespread use for many years. Future empirical testing of parameter settings should be done but with a broader test set of more contemporary drug safety issues during real-world use of medicinal products, as well as attempt to demonstrate the timeliness of potential issue identification.

Converting raw THIN to CDM THIN is a resource-intensive project. It took approximately two full-time person-years to complete this database conversion and validation, which is similar to the average time spent converting a database into the OMOP CDM for OMOP-distributed data partners [12]. Strong SAS programming, database management and statistical skills, as well as expertise with the THIN database, were crucial for this conversion. Several important limitations of CDM THIN should be noted: (i) incomplete mapping of medical, drug and laboratory codes in raw THIN into the standard terminology dictionary; (ii) information loss due to THIN and CDM data structure differences; for example, comments (text field) from GPs in raw THIN were not mapped, which may have impact for a detailed analysis using the diagnosis or prescription data embedded in the comments; (iii) defining appropriate assumptions without analyses defined a priori can be problematic; for example, assumptions of condition ending date being the same as condition starting date will clearly not be appropriate for all conditions; and (iv) significant IT resource commitments and challenge in updating the CDM tables. In addition, some analysis methods are computationally intensive and require large storage space; thus, how to ensure timely updates of CDM THIN based on the updated source database for a rapid assessment can be particularly challenging for routine active surveillance purposes.

We have shown that THIN can be implemented in a CDM, specifically the OMOP CDM. Clearly, THIN could potentially be implemented in other CDMs. The complexity of the OMOP CDM structure (e.g. multiple data and reference tables and standard terminology dictionary), while enhancing its uses also increases difficulty with

implementation and the need for standardized tools to ensure validity and transparency of the implementation. While the OMOP has shown that such tools can be developed and used to test the effectiveness of CDM implementations for research, one could see an argument that a simpler CDM might be preferred for ease of use in specific applications. Further work is needed to determine which CDM is most useful for THIN and ongoing work is needed to compare CDMs with the vision of consolidating to one or a few (e.g. one for claims, one for EHRs) CDMs for widespread international use of heterogeneous observational datasets. The value of a CDM in active surveillance is reliant on common use of the CDM and thus efforts need to be made both to implement databases into the CDM structure and also to maintain high-quality mapping of the database over time. Processes to ensure transparent routine processes for updating any CDM and related libraries to minimize the risk of diverging versions of CDMs (potentially on the same underlying raw data) will be a necessity to maximize the values of CDMs. The OMOP CDM can provide an important ongoing tool for analyses in real-world data, with appropriate processes and training to ensure continual harmonized use of the CDM by different users and organizations.

5 Conclusions

This study demonstrated that converting THIN data into the OMOP CDM was feasible though imperfect. DOIs and HOIs predefined by the OMOP in this study were constructed with minimal loss of information and can be used for active surveillance methodological research. CDM THIN allows the application of three standard analysis methods developed by the OMOP. It offers the potential as a valuable tool for multiple aspects of pharmacoepidemiological research, particularly if the unique features of UK EHRs are more fully incorporated in the OMOP library. Standardizing database structure in a distributed system of databases through a CDM has practical benefits. However, there is limited research on whether such benefits come at the risk of information loss reducing the value of the database for pharmacoepidemiological research. This study highlights the need for more research into the strengths and limitations of the use of CDMs to inform appropriate effective routine, real-world surveillance.

Acknowledgements We would like to thank OMOP for designing a CDM and associated tools and algorithms, and for making the programs and associated detailed documentation readily available. In particular, we would like to thank Drs Patrick Ryan and Christian Reich for their advice on the implementation of the THIN database into the OMOP CDM format. We would also like to thank Dr. Manfred Hauben for his review and specific advice to the mapping of the THIN database into the OMOP CDM.

Funding and Conflicts of Interest No sources of funding were used to conduct this study or prepare this manuscript. Xiaofeng Zhou, Sundaresan Murugesan, Qing Liu, Bing Cai and Andrew Bate are full-time employees of Pfizer and hold Pfizer stock options/stocks. Harshvinder Bhullar is an employee of CSD, the company that supplies the THIN database. Chuck Wentworth was a part-time contract employee at Pfizer during the time the study was conducted.

References

1. Strom BL. Pharmacoepidemiology. 4th ed. West Sussex: John Wiley & Sons Ltd; 2005.
2. Hall GC, Sauer B, Bourke A, et al. Guidelines for good database selection and use in pharmacoepidemiology research. Pharmacoepidemio Drug Saf. 2012;21(1):1–10.
3. Arnold RJ, Balu S. Retrospective database analysis. In: Arnold RJG, editor. Pharmacoeconomics: from theory to practice. Boca Raton: CRC Press; 2009. p. 59–82.
4. Foundation for National Institutes of Health. Observational Medical Outcomes Partnership (online). http://omop.fnih.org/node/22. Accessed 1 May 2010.
5. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med. 2010;153(9):600–6.
6. Coloma PM, Schumie MJ, Trifiro G, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepdiemio Drug Saf. 2011;20(1):1–11.
7. Maro JC, Platt R, Holmes JH, et al. Design of a national distributed health data network. Ann Intern Med. 2009;151(5):341–4.
8. Lewis JD, Schinnar R, Bilker W, et al. Validation studies of the health improvement network(THIN) database for pharmacoepidemiology research. Pharmacoepdiemio Drug Saf. 2007;16:393–401.
9. Smeeth L, Cook C, Thomas S, et al. Risk of deep vein thrombosis and pulmonary embolism after acute infection in a community setting. Lancet. 2006;367(9516):1075–9.
10. 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):171–7.
11. Gonzalez EL, Johansson S, Wallander MA, et al. Trends in the prevalence and incidence of diabetes in the UK: 1996–2005. J Epidemiol Community Health. 2009;63(4):332–6.
12. Overhage JM, Ryan PB, Reich CG, et al. Validation of a common data model for active surveillance research. J Am Med Inform Assoc. 2012;19(1):54–60.
13. Reisinger S, Ryan PB, O'Hara DJ, et al. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. J Am Med Inform Assoc. 2010;17:652–62.
14. Van Le H, Beach KJ, Powell G, et al. Performance of a semi-automated approach for risk estimation using a common data model for longitudinal health databases. Stat Methods Med Res. Epub 2011 Jun 16.
15. Foundation for National Institutes of Health. OMOP Common Data Model (CDM) specifications, version 2.0, November 3, 2009 (online). http://omop.fnih.org/ETLProcess. Accessed 2 May 2010.
16. Cegedim Strategic Data (CSD) Medical Research Group, UK. THIN data content (online). http://csdmruk.cegedim.com/our-data/data-content.shtml. Accessed 3 May 2010.
17. Ryan PB. Establishing a condition era persistence window for active surveillance, Jan 12, 2010 (online). http://omop.fnih.org/OMOPWhitePapers. Accessed 4 May 2010.

18. Ryan P, OMOP PI, Establishing a Drug Era Persistence Window for Active Surveillance, Jan 25, 2010 (online). http://omop.fnih.org/OMOPWhitePapers. Accessed 4 May 2010.
19. Ryan PB, Reich C, Welebob E. Managing data quality for an active surveillance system (online). http://omop.fnih.org. Accessed 12 May 2010.
20. Foundation for National Institutes of Health, OMOP. OSCAR – Observational Source Characteristics Analysis Report (OSCAR) design specification and feasibility assessment (online). http://omop.fnih.org/OSCAR. Accessed 5 May 2010.
21. Foundation for National Institutes of Health, OMOP. NATHAN – Utility of Natural History Information (online). http://omop.fnih.org/NATHAN. Accessed 6 May 2010.
22. Foundation for National Institutes of Health, OMOP. Generalized Review of OSCAR Unified Checking (online). http://omop.fnih.org/GROUCH. Accessed 1 May 2010.
23. OMOP. Specifications for implementation of standard vocabularies in observational data analysis, version 3.0, June 2010 (online). http://omop.fnih.org/vocabularies. Accessed 4 May 2010.
24. Zorych I, Madigan D, Ryan P, et al. Disproportionality methods for pharmacovigilance in longitudinal observational databases. Stat Methods Med Res. Epub 2011 Aug 30.
25. Bate A, Evans SJW. Quantitative signal detection using spontaneous ADR reporting. Pharmacoepidemio Drug Saf. 2009;18(6):427–36.
26. Ryan PB, Madigan D, Method performance results from the Health Outcomes of Interest Experiment: part 1 (online). http://omop.fnih.org/OMOP2011Symposium. Accessed 5 May 2010.
27. Hubbard R, Farrington P, Smith C, et al. Exposure to tricyclic & selective serotonin reuptake inhibitor antidepressants and the risk of hip fracture. Am J Epidemiol. 2003;158:77–84.
28. Maclure M, Fireman B, et al. When should case-only designs be used for safety monitoring of medical products? Pharmacoepidemiol Drug Saf. 2012;21(Suppl. 1):50–61.
29. Schneeweiss S, Rassen JA, Glynn RJ. et al High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology. 2009;20:512–22.
30. Rassen JA, Schneeweiss S. Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system. Pharmacoepidemiol Drug Saf. 2012;21(Suppl. 1):41–9.
31. Whitaker HJ, et al. The methodology of self-controlled case series studies. Stat Methods Med Res. 2009;18(1):7–26.
32. Foundation for National Institutes of Health, OMOP. Observational analysis methods and methods library (online). http://omop.fnih.org/MethodsLibrary. Accessed 12 May 2010.
33. Cegedim Strategic Data (CSD) Medical Research Group, UK. THIN bibliography (online). http://csdmruk.cegedim.com/bibliography/bibliography_01.shtml. Accessed 10 May 2010.
34. Hardoon SL, Whincup PH, Petersen I, et al. Trends in longer-term survival following an acute myocardial infarction and prescribing of evidenced-based medications in primary care in the UK from 1991: a longitudinal population-based study. J Epidemiol Community Health. 2011;65(9):770–4.
35. van Puijenbroek EP, Bate A, Leufkens HG, et al. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf. 2002;11(1):3–10.
36. Bate A, Evans SJW. Quantitative signal detection using spontaneous ADR reporting. Pharmacoepidemiol Drug Saf. 2009;18(6):427–36.
37. Norén NG, Hopstadius J, Bate A, et al. Temporal pattern discovery in longitudinal electronic patient records. Data Min Knowl Discov. 2010;20(3):361–87.
38. Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. Pharmacoepidemiol Drug Saf. 2010;19:858–68.
39. Ryan PB. Defining a reference set for evaluating the performance of active surveillance methods. 31 January 2010 (online). http://omop.fnih.org/OMOPWhitePapers. Accessed 4 May 2010.