Journal of Biomedical Informatics 52 (2014) 364–372

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Integrating electronic health record information to support integrated care: Practical application of ontologies to improve the accuracy of diabetes disease registers

Siaw-Teng Liaw [a,b,c,*], Jane Taggart [b], Hairong Yu [b], Simon de Lusignan [d], Craig Kuziemsky [e], Andrew Hayen [a]

[a] *School of Public Health and Community Medicine, UNSW Medicine, Sydney, Australia*
[b] *Centre for PHC & Equity, UNSW Medicine, Sydney, Australia*
[c] *Academic General Practice Unit, South Western Sydney Local Health District, NSW, Australia*
[d] *University of Surrey, Guildford, UK*
[e] *Telfer School of Management, University of Ottawa, Ottawa, Canada*

ARTICLE INFO

Article history:
Received 17 March 2014
Accepted 23 July 2014
Available online 1 August 2014

Keywords:
Case finding
Integration
Patient register
EHR
Ontology
Knowledge engineering

ABSTRACT

Background: Information in Electronic Health Records (EHRs) are being promoted for use in clinical decision support, patient registers, measurement and improvement of integration and quality of care, and translational research. To do this EHR-derived data product creators need to logically integrate patient data with information and knowledge from diverse sources and contexts.
Objective: To examine the accuracy of an ontological multi-attribute approach to create a Type 2 Diabetes Mellitus (T2DM) register to support integrated care.
Methods: Guided by Australian best practice guidelines, the T2DM diagnosis and management ontology was conceptualized, contextualized and validated by clinicians; it was then specified, formalized and implemented. The algorithm was standardized against the domain ontology in SNOMED CT-AU. Accuracy of the implementation was measured in 4 datasets of varying sizes (927–12,057 patients) and an integrated dataset (23,793 patients). Results were cross-checked with sensitivity and specificity calculated with 95% confidence intervals.
Results: Incrementally integrating Reason for Visit (RFV), medication (Rx), and pathology in the algorithm identified nearly100% of T2DM cases. Incrementally integrating the four datasets improved accuracy; controlling for sample size, data incompleteness and duplicates. Manual validation confirmed the accuracy of the algorithm.
Conclusion: Integrating multiple data elements within an EHR using ontology-based case-finding algorithms can improve the accuracy of the diagnosis and compensate for suboptimal data quality, and hence creating a dataset that is more fit-for-purpose. This clinical and pragmatic application of ontologies to EHR data improves the integration of data and the potential for better use of data to improve the quality of care.

1. Introduction

Electronic Health Records (EHRs) and informatics-enabled integrated care can improve chronic disease management (CDM). There are benefits for health care providers and consumers through more accurate and timely information exchange, improved work efficiency by avoiding repetition of information collection and tests, and better decision-making [1]. A widely used CDM is the Chronic Care Model (CCM) [2,3]. The CCM has six dimensions: health care organization, delivery system design, decision support, clinical information systems, self-management support, and community resources/policies to optimize integrated CDM. The CCM describes an activated patient engaged with an activated health care team to optimize patient-centred care [4] and the activated patients involved in self-management.

Patient registers or lists derived from routinely collected data in EHRs may be developed through “phenotyping” [5] or “case-finding” algorithms to identify cases. The increasing use of health-related social media, particularly in socially shaped diseases such as obesity and depression, can and should be exploited to

[*] Corresponding author. Address: General Practice Unit, Fairfield Hospital, PO Box 5, Fairfield, NSW 1860, Australia. Fax: +61 2 96168400.
E-mail address: siaw@unsw.edu.au (S.-T. Liaw).

http://dx.doi.org/10.1016/j.jbi.2014.07.016
1532-0464/© 2014 Elsevier Inc. All rights reserved.

enhance traditional ways of using EHRs to support care and care delivery [6].

The limitations of traditional EHR-based registers have been reported in the USA [7], UK [8,9], Australia [10] and Canada [11]. The Canadian diabetes registry to support patient identification and disease tracking was scrapped due to cost overruns and failure to deliver to expectations. Basic questions about the accuracy and completeness of EHR-based registers were only partially answered. The sensitivity (extent to which all patients with the disease were included) and specificity (extent to which patients who do not have the disease were excluded) of these registers were often not recorded.

The STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) and REporting of studies Conducted using Observational Routinely collected Data (RECORD) initiatives recognize this gap. However, RECORD and STROBE are in their infancy and, in our view, lack sufficient and necessary support of clinical informatics [12,13] and knowledge engineering communities. There is scope to make greater use of semantic web technologies to exploit propositional knowledge (http://www.phekb.org/) as well as other domain knowledge: about health professionals, patients' health status, and the health system.

The design of many EHRs and EHR-based disease registers is not transparent [14], mostly to protect intellectual property. The quality of disease registers has been critiqued in the UK [9,14,15] and those within an electronic Practice Based Research Network (ePBRN) in Australia [16,17]. Whilst some progress has been made, the core outstanding issue is uncertainty about true negatives, namely people without diabetes who are identified as such.

1.1. Theoretical approach

Our theoretical approach combines ontological and realist perspectives. We use Gruber's classic description of an ontology as the "specification of a conceptualization" [18]. We incorporate Carlsson's work as the guiding principle within a critical realist approach; [19] whilst adopting the practicality of Pawson and Tilley's method of conducting a realist evaluation. They describe this straightforwardly as: "What works for whom in what context" [20]. Pawson and Tilley also describe how: "Context + Mechanism = Impact/Outcome (CMIO)." [21]. Finally we take a knowledge engineering approach (KE) to linking key data, information and knowledge.

1.1.1. Ontology

We develop a phenotype ontology for type 2 diabetes mellitus (T2DM) with a focus on supporting multidisciplinary integrated care of patients with chronic diseases capable of supporting the implementation of the CCM. Our ontological approach recognizes that the quality of EHRs and EHR-derived utilities such as patient registers is influenced by the data, knowledge modeling, system architecture, implementation protocols, training and support and associated knowledge management and information governance processes. We adopt an ontological layered approach [22] that incorporated rule-based methods, clinical guidelines and data quality dimensions.

Ontologies can potentially support several categories of integration including:

- *Data integration*: from disparate clinical data sources within and across EHRs.
- *Knowledge integration*: from diverse health and social professionals' knowledge.
- *Clinical integration*: linking clinical concepts to model the phenotype.
- *Interdisciplinary integration*: bringing together multiple disciplines to model and support multidisciplinary coordinated care and information exchange in CDM over the patient's journey through a complex ecosystem of clinical and social factors and contexts such as co-morbidities, health risks, health financing and insurance.

We used the Australian extension to the Systematized Nomenclature Of Medicine Clinical Terms (SNOMED CT-AU®) as a standardized terminology for knowledge representation in multidisciplinary clinical practice. SNOMED CT-AU® is an ontology which formally defines classes of medical procedure, pharmaceutical or biologic product, and body structure (http://www.nehta.gov.au/our-work/clinical-terminology/snomed-clinical-terms). In Australia, SNOMED CT-AU® Ontology (SCAO) is available in Web Ontology Language (OWL) from the National E-Health Transition Authority (NEHTA). Preliminary tests of SNOMED CT-AU demonstrated its suitability for integration with our specified ontology to identify patients with T2DM.

1.1.2. Realist approach

Our realist CMIO framework is applied as follows:

- The context is the health system, including multidisciplinary teams, continuity and integration of care, quality improvement indicators, disease surveillance and population health. We emphasize a need to collect and manage complete, correct, consistent and timely information about the cycle of care, risk factors, disease indicators, quality of life and patient satisfaction.
- The mechanism to meet this need are the informatics and knowledge engineering tools, including the EHR, its data and ontologies to conceptualize the concepts and relationships required to implement and evaluate evidence-based best practice guidelines for a range of contexts and purposes.
- The outcome in this study is the accuracy of the disease register.

We aim to embed "ontological" thinking in clinical practice to facilitate more realistic and relevant translation of clinical practice into knowledge modeling to improve knowledge collection, management and use in practice as well as the design of relevant decision support systems. The congruence of clinical and technical ontologies is essential to facilitate semantic and syntactic integration to promote the development of more proactive and intelligent CDM systems to reduce the burden on providers while still integrating patient information to guide integrated multidisciplinary practice, research and policy.

1.1.3. Knowledge engineering (KE)

A KE approach [23] has the potential to integrate knowledge with the complex processes, sophisticated functions and rich information inherent in CDM. Ontologies are especially important when dealing with people with multiple chronic diseases, whose health data are often distributed among different health and social care providers and in different formats.

1.2. Study setting

The ePBRN in South West Sydney provided the dataset to create an EHR-based register of patients with diabetes, to assess the accuracy of the tools and ontology-specified algorithm, and to test the resulting informatics and KE infrastructure developed. The records were linked, using a probabilistic matching and record linkage tool [24], to assess the extent of duplicate patient records and to avoid double counting of patients who used multiple practices in the study area (medical neighborhood).

This paper describes the use of ontology and KE tools to design, develop, implement and validate the T2DM phenotyping algorithm component of a comprehensive ontological approach to identifying and managing patients with T2DM from primary care EHRs.

2. Methods

We adopted an approach we had previously piloted [25]:

2.1. Specification and conceptualization

The diabetes conceptual framework and, subsequent, multi-attribute T2DM phenotyping algorithm was developed using existing clinical knowledge and best practice guidelines [26,27], and our research into guideline implementation research findings [17].

A unified context was specified to act in the presence of differences in terminology and semantics from different EHRs, support reusability and integration of data. This approach also supported the development of automated systems for data annotation, extraction and linkage, information retrieval, data quality management (DQM) and natural-language processing [28]. By incorporating defined rules, the ontology generated logical inferences and consequently controlled the inclusion/exclusion of relevant objects [29]. Examples of this include: the patient with T2DM-related Reason for Visit (RFV), medication (Rx), specific pathology test (e.g. glycated haemoglobin (HbA1c), plasma glucose), service (e.g. referral to a diabetologist), benefits paid (e.g. T2DM cycle of care), risk factors (e.g. body mass index (BMI)), and/or a Family History (FH) [30].

The conceptualization of the ontology was finalized through an iterative process with the clinicians participating in the electronic Practice Based Research Network (*e*PBRN).

2.2. Formalization and implementation

The ePBRN pilot dataset is an aggregate of EHR data from four participating practices. These practices contain 93,000 pseudonymised patients, with over a 13 year review period more than 1.6 million consultation records, 690,000 diagnoses (including RFV where it implies a diagnosis), 1 million prescription records, and 1.6 million pathology records. We used Microsoft SQL Server (http://www.microsoft.com/en-us/sqlserver/default.aspx) to implement the algorithm. The repository for the ePBRN data was created with five indexed tables:

- Diagnosis_Table including RFV,
- Prescription_Table for Rx,
- Pathology_Table for laboratory test name and result,
- Measures_Table provided additional information (e.g. BMI) and the
- Family_History_Table.

We used these tables to support decisions about the patients who do not have a DM label but have some relevant information that suggest that they may have DM. This hybrid structure of relational and non-relational database attributes was developed to deliver maximum efficiency and flexibility. The significant computational cost of a query across a number of large tables is reduced by indexing and restricting the number of tables to those listed above. Retrieval time is proportional to log(n), where n is the number of records. A properly cached index meant that lookup from our 1.6 million-row table was done in milliseconds. The algorithm was implemented in T-SQL (Transact-SQL) (http://msdn.microsoft.com/en-us/library/ms189826%28v=sql.90%29.aspx), an extension of SQL (Structured Query Language) supported by Microsoft. Fig. 1 summarizes the flow of the T2DM case finding algorithm.

Finally, we checked for and excluded duplicate or multiple records, to reduce the likelihood of double-counting in a geographical area where patients may use a number of general practices and hospital-based services. While the duplicate check can be done at every phase, we chose to check at the end to avoid excluding DM-related data that may exist separately in the duplicated records, either within a single practice or across a number of practices.

2.3. Architecture of ontology infrastructure

The architecture to facilitate the classification and retrieval of DM patients is shown in Fig. 2.

Patient data were extracted and selected from individual participating general practice EHRs by GRHANITE™ (http://www.grhanite.com), which provides a user accessible data repository sitting over a secure server [24]. Patient data, associated with ontology classes or properties, are sorted using -ontopPro- (http://ontop.inf.unibz.it/), which is theoretically based on the Ontology-Based Data Access principles [31]. The knowledge component of the infrastructure, related to conceptual terminologies was defined by the specification ontology, and built using SNOMED CT-AU and Web Ontology Language (OWL) (http://www.w3.org/TR/owl-features/) using the Protege open source ontology editor (http://protege.stanford.edu/).

We used hierarchical conceptual modeling to create the T2DM phenotype ontology, guided by the Australian National Guidelines for T2DM and discussions with the research team and general practitioners participating in the ePBRN. The formalized T2DM phenotype ontology consists of 4 main classes: Actor, Content, Mechanism and Impact; and 68 subclasses with object or data properties specified for the study objective. We modeled the domain specific knowledge for T2DM identification in SNOMED CT-AU Ontology (SCAO), which has more than 300,000 concepts.

We used T-SQL™ to link the server objects in the SQL Server to the heterogeneous datasets. The SQL query results were mapped, using -ontopPro-, to the patient data in the T2DM phenotype ontology; these in turn were associated with relevant classes. This meant that the schematic or semantic heterogeneity challenges faced were solved at either data or ontology level. The mapping mechanism supplied by -ontopPro- enabled: the populating of class members, assigning of property values, and incorporation of schematic data in the ePBRN repository with semantic concepts provided by the ontologies. Finally, we used T-SQL™ to implement the phenotyping algorithm on the knowledge engineered ontology infrastructure to identify patients with T2DM.

2.4. Testing and validating the accuracy of the implementation of the algorithm

2.4.1. Measures of accuracy

We adopted the International Standard Organization (ISO) 5725-1 definition of *accuracy* as consisting of *trueness* (proximity of measurement results to the true value) and *precision* (repeatability or reproducibility of the measurement) (see Fig. 3). Cases of DM in medical records can be true or false positives (TP or FP), true or false negatives (TN, FN). *Accuracy* is the proportion of true results (TP + TN) in the population. *Positive predictive value* (*PPV*) is defined as all positive test results (both TP + FP) that are truly positive. *Sensitivity* (also called *true positive rate*) measures the proportion of actual positives that are correctly identified as such (TP/(TP + FP)). *Specificity* (also called *true negative rate*) measures the proportion of negatives that are correctly identified as such

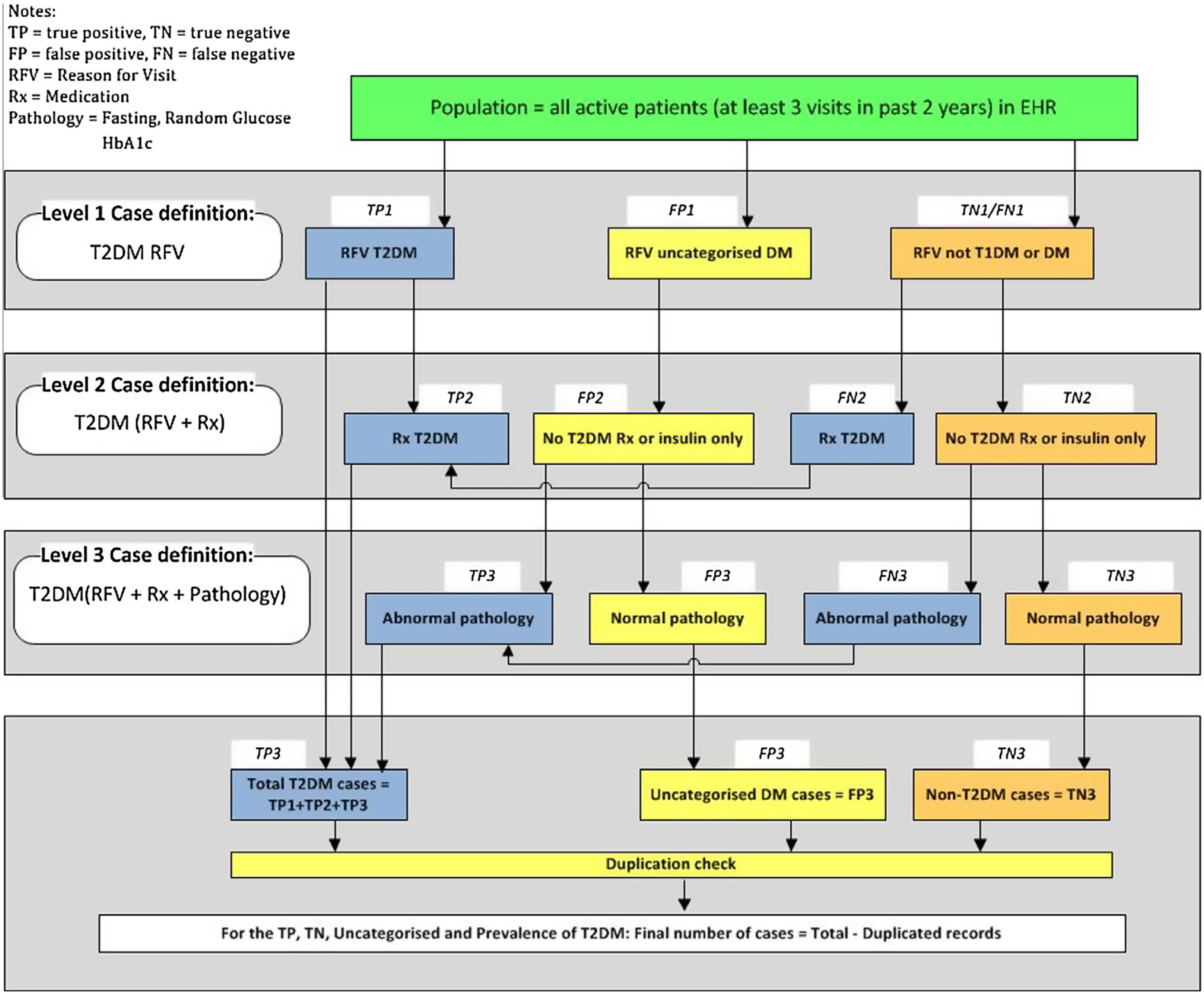

Notes:
TP = true positive, TN = true negative
FP = false positive, FN = false negative
RFV = Reason for Visit
Rx = Medication
Pathology = Fasting, Random Glucose
HbA1c
Population = all active patients (at least 3 visits in past 2 years) in EHR
Level 1 Case definition:
T2DM RFV
TP1
RFV T2DM
FP1
RFV uncategorised DM
TN1/FN1
RFV not T1DM or DM
Level 2 Case definition:
T2DM (RFV + Rx)
TP2
Rx T2DM
FP2
No T2DM Rx or insulin only
FN2
Rx T2DM
TN2
No T2DM Rx or insulin only
Level 3 Case definition:
T2DM(RFV + Rx + Pathology)
TP3
Abnormal pathology
FP3
Normal pathology
FN3
Abnormal pathology
TN3
Normal pathology
TP3
Total T2DM cases = TP1+TP2+TP3
FP3
Uncategorised DM cases = FP3
TN3
Non-T2DM cases = TN3
Duplication check
For the TP, TN, Uncategorised and Prevalence of T2DM: Final number of cases = Total - Duplicated records

Fig. 1. Flowchart of the T2DM case finding algorithm.

(TN/(TN + FN)). If prevalence is known, *PPV* may be determined from *Sensitivity* (*Sn*) and *Specificity* (*Sp*), using the equation:

$$PPV = Sn * prevalence/(Sn * prevalence + (1 - Sp) * (1 - prevalence)).$$

From the *Sn*, *Sp* and *PPV* (precision), we can compute the *F*-Measure, weighted harmonic mean that takes into account both the precision and recall of a test. The receiver operating characteristic (ROC) curve is a graphical demonstration of the relationship between sensitivity and specificity. We are only using a binary classification algorithm and therefore not varying the decision threshold. In the case of a binary test, the area under the ROC curve can be shown, using simple geometry, to be [(*Sn* + *Sp*)/2].

Sample size: To achieve 95% confidence intervals (CI) for a conservative estimate of 0.5 (+/−0.025) for *Sn*/*Sp*, for single or multiple attributes, we calculated the sample size to be 5000 patients (Australian National Statistical Services online sample size calculator; http://www.nss.gov.au/nss/home.nsf/pages/Sample+size+calculator?OpenDocument).

2.4.2. Testing process

The T2DM phenotyping algorithm was first tested in the smallest ePBRN practice (Practice #1), with the other practice datasets of increasing sizes added one at a time. This enabled an assessment of any effects of combining practice datasets of varying data quality [32], as well as addressing semantic and technical integration issues. Previous studies have found more than 300 terms which might be associated with a diagnosis of DM [16].

The first pass through the data sets used Reason for Visit (RFV) as the filter (Fig. 1).

In the second pass, a DM medication (Rx) filter was used on all FP (Diagnoses/RFV of DM). All FP cases with a DM medication (both oral and insulin) were re-categorized as TP (based on Diagnosis/RFV or Rx). Presence of insulin only meant that the case was T1DM and therefore not T2DM and therefore a TN. The *Sn*/*Sp* (*for diagnosis RFV or Rx*) was calculated (see Fig. 3).

The third pass identified diagnostic pathology (Path) tests for DM (HbA1c ⩾ 6.5%, fasting plasma glucose ⩾ 7, random plasma Glucose ⩾ 11.1) as the filter. Abnormal DM specific tests meant re-categorization as TP (*RFV or Rx or Path*). The *Sn*/*Sp* (*diagnosis, RFV or Rx or Path*) was calculated.

The fourth pass used the presence of a DM-related diagnosis/RFV on at least two visits, a family history of diabetes and BMI ⩾ 30 as filters to gain further information about patients who may be unrecognized diabetics or have pre-diabetes. We also calculated 95% confidence intervals for the *Sn*/*Sp* using the Wilson score method without continuity correction [33].

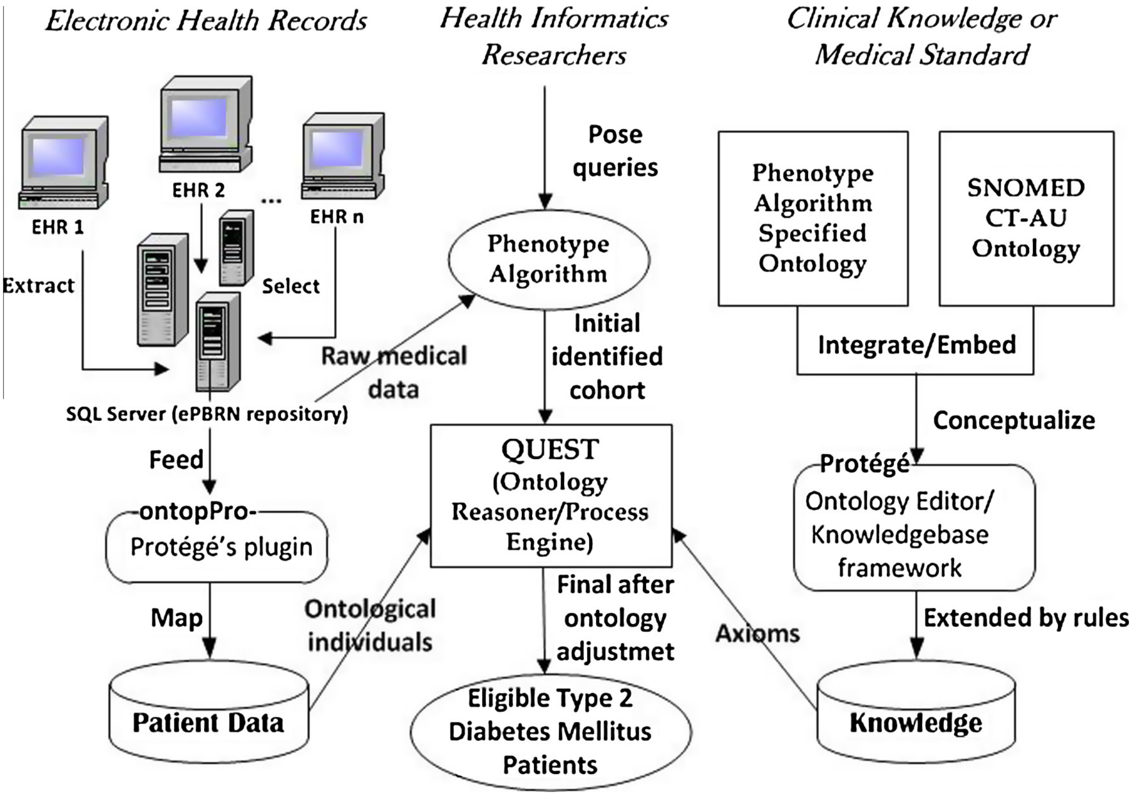

Electronic Health Records
Health Informatics Researchers
Clinical Knowledge or Medical Standard
EHR 1
EHR 2
...
EHR n
Extract
Select
SQL Server (ePBRN repository)
Pose queries
Phenotype Algorithm
Raw medical data
Initial identified cohort
Phenotype Algorithm Specified Ontology
SNOMED CT-AU Ontology
Integrate/Embed
Conceptualize
Feed
ontopPro-
Protégé's plugin
QUEST (Ontology Reasoner/Process Engine)
Protégé
Ontology Editor/ Knowledgebase framework
Map
Ontological individuals
Final after ontology adjustmet
Axioms
Extended by rules
Patient Data
Eligible Type 2 Diabetes Mellitus Patients
Knowledge

Fig. 2. Ontology architecture to classify and retrieve patients with T2DM.

2.4.3. Triangulation

The accuracy (trueness + precision) was triangulated by feedback from clinicians at Practice 1 about the list of T2DM cases identified. We did not obtain feedback on the list of negatives.

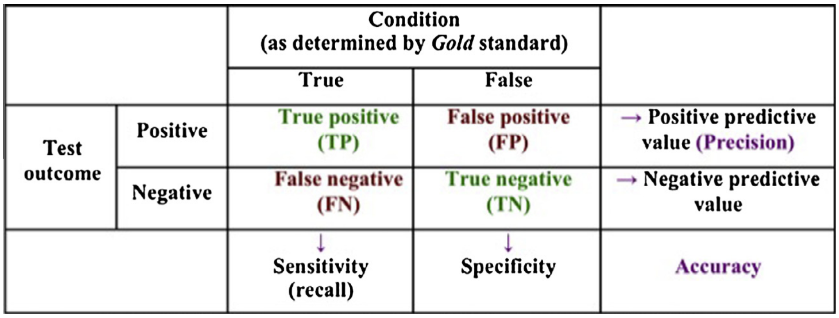
3. Results

T2DM patients identified at each stage of the algorithm for the ePBRN data repository ($N = 23{,}793$) were sorted into TP, TN, FP and FN groups (Fig. 4). Similar summary figures were created for each source practice datasets. We validated the accuracy of the T2DM cases identified in Practice #1.

The *Sn*/*Sp* of the algorithm and 95% confidence intervals (CI) and completeness of data in the patient record were calculated for each of the datasets and summarized in Tables 1 and 2 respectively. Table 1 demonstrates that, when taken together, the RFV and Rx will accurately identify around 95% of the cases; when abnormal pathology was included, the accuracy was almost unity. Table 1 also shows the increasing precision (decreasing confidence intervals) as the size of the source practice data set increases down the table to the total of all 4 practices. The precision is adequate once the data set meets or exceeds a required sample size.

While we conservatively calculated the required sample size as 5000 patients, it is possible that 3500 would be adequate (Table 2). The size of the data set influenced the PPV of the algorithm, without appreciable depreciation in the Sp or Negative Predictive Value (NPV).

The result for Practice 1 ($n = 927$) reflected the participating GPs' reports that they did not routinely record diabetes type in the RFV field, which affirmed the value of this approach as the cases would not be identified as T2DM if we relied on the diagnosis/RFV alone. The large 95% confidence intervals reflected the small size of Practice 1 and, perhaps, the completeness of the dataset (Table 3). The range of estimates of prevalence of T2DM ranged from 2.6% in Practice 1 to 9.3% in Practice 2; the pattern was not consistently related to the size of the data set, suggesting



		Condition (as determined by *Gold* standard)		
		True	False	
Test outcome	Positive	True positive (TP)	False positive (FP)	→ Positive predictive value (Precision)
	Negative	False negative (FN)	True negative (TN)	→ Negative predictive value
		↓ Sensitivity (recall)	↓ Specificity	Accuracy

Fig. 3. Contingency table summarizing accuracy measures adopted.

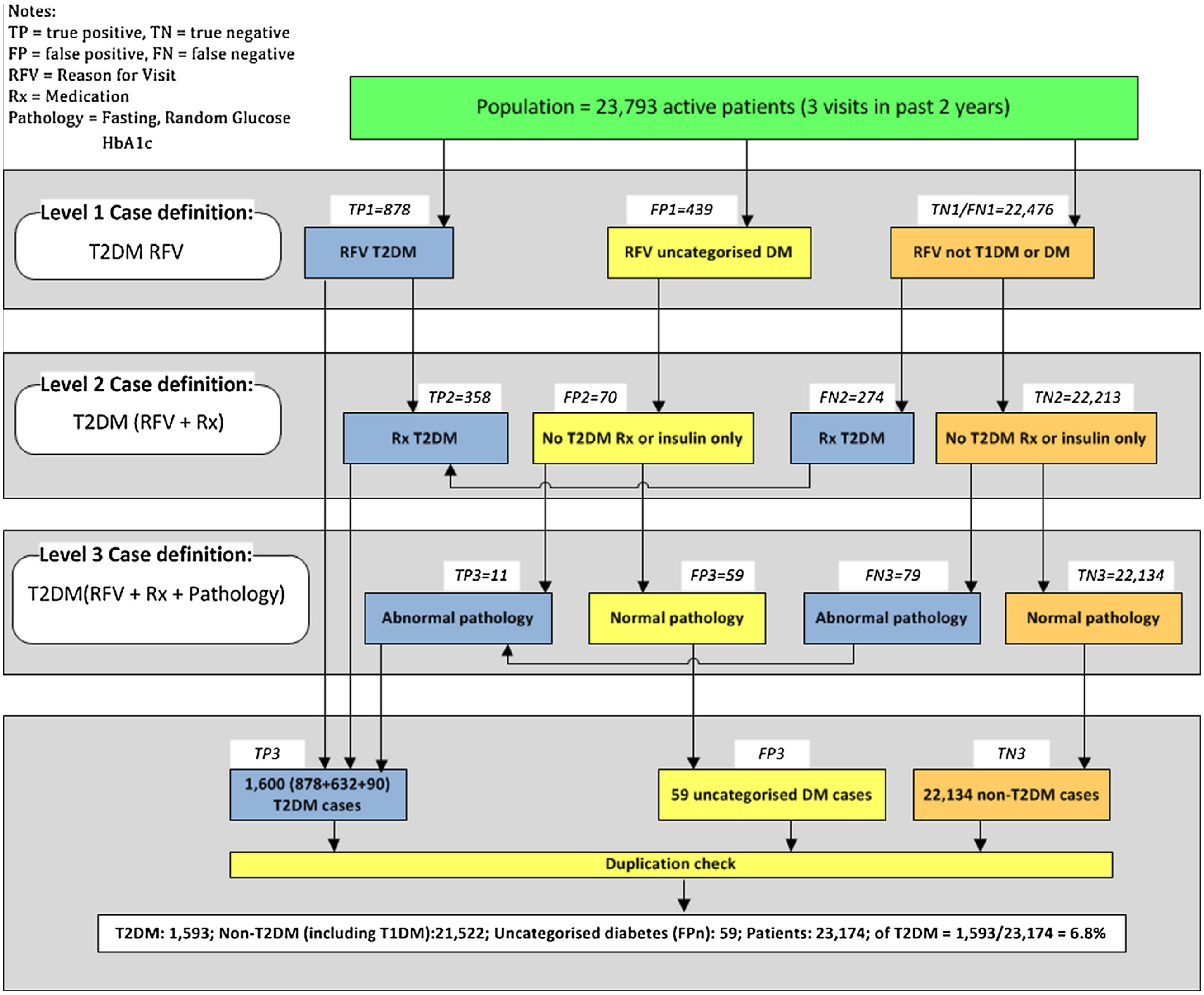

Notes:
TP = true positive, TN = true negative
FP = false positive, FN = false negative
RFV = Reason for Visit
Rx = Medication
Pathology = Fasting, Random Glucose
HbA1c
Population = 23,793 active patients (3 visits in past 2 years)
Level 1 Case definition:
T2DM RFV
TP1=878
RFV T2DM
FP1=439
RFV uncategorised DM
TN1/FN1=22,476
RFV not T1DM or DM
Level 2 Case definition:
T2DM (RFV + Rx)
TP2=358
Rx T2DM
FP2=70
No T2DM Rx or insulin only
FN2=274
Rx T2DM
TN2=22,213
No T2DM Rx or insulin only
Level 3 Case definition:
T2DM(RFV + Rx + Pathology)
TP3=11
Abnormal pathology
FP3=59
Normal pathology
FN3=79
Abnormal pathology
TN3=22,134
Normal pathology
TP3
1,600 (878+632+90) T2DM cases
FP3
59 uncategorised DM cases
TN3
22,134 non-T2DM cases
Duplication check
T2DM: 1,593; Non-T2DM (including T1DM):21,522; Uncategorised diabetes (FPn): 59; Patients: 23,174; of T2DM = 1,593/23,174 = 6.8%

Fig. 4. T2DM cases identified by algorithm to 3rd level in ePBRN data repository.

that data completeness was a factor. The prevalence of T2DM in the combined data set was 6.7%.

4. Discussion

By integrating diagnosis/RFV, Rx and pathology, this algorithm could identify T2DM cases with an accuracy of 99.9% and PPV of 96.4%. Using more than the three data elements did not improve the accuracy significantly. The 6.7% prevalence of T2DM calculated by the ontology-based "*RFV or Rx or Pathology*" algorithm was comparable with the prevalence of 6.5% reported by the local health authority. In comparison, the prevalence of the T2DM reported using RFV alone was only 5.8%. The 1:20 ratio of T1DM to T2DM in the ePBRN data set was consistent with prevalence estimates in the literature [27,34]. The size of the data repository (calculated practice list size of 5000) was adequate for phenotyping; although the actual requirement may be less. This confirms the importance of calculating and achieving an adequate sample size to answer specific questions.

The increasing yield of cases identified by using three attributes (diagnosis/RFV, medication and pathology), while maintaining the accuracy, suggested that this ontological multi-attribute approach was more accurate than just relying on single attributes such as disease codes. It provided some insight into the reliability and validity of routinely collected data in the source practice records, and likely degree of inter-practice variation. The method is transparent and can be built on, with feedback given to practices and practitioners to improve data recording and completeness. The other dimensions of data quality (DQ) that the ePBRN research is focused on – correctness and consistency [16,17] – have been addressed within the ontology.

The findings of this study can guide the development of thresholds for other possible associations and/or predictors of T2DM and for other chronic diseases or co-morbidity, BMI (obesity), ethnicity, and gender. The ontological approach has the potential to help preserve the fidelity of searches in the real world of clinical practice where the patient often requires multidisciplinary integrated care for multiple health issues related to multimordibity and polypharmacy. This study validates the ontology-based approach to integrate the attributes of a diabetic patient at the data (diagnosis/RFV) and knowledge (diagnosis/RFV, Rx and Path) levels. We have conducted but not reported on analyses on integration at the clinical (BMI + FH) and interdisciplinary (referrals and use of T2DM related services provided by other service providers) levels as they were superfluous to this exercise to examine the accuracy of the T2DM phenotyping algorithm. A substantive evaluation of the inter-disciplinary aspect of this approach is planned for future studies, using another clinical diagnosis.

Table 1
Increasing *Sn*/*Sp* with increasing size of source practice and ePBRN data repository.

Diabetes attribute	TP	FP	TN	FN	*Sn*%	95% CI	*Sp*%	95% CI
Source Practice 1 (n = 927)⁎ [T1DM = 2 (identified at RFV)]								
Reason for Visit (RFV)	2	36	867	22	8.3	2.3–25.8	96.0	94.5–97.1
RFV or medication (Rx)	21	20	883	3	87.5	69.0–95.6	97.8	96.6–98.6
RFV or Rx or pathology	24	18	885	0	100	86.20–100	98.0	96.9–98.7
After duplicates removed	24	18	885	0	100	86.20–100	98.0	96.9–98.7
Source Practice 2 (n = 3,699)⁎ [T1DM = 7 (4 identified at RFV and 3 at Rx)]								
Reason for Visit (RFV)	44	161	3196	298	12.9	9.7–16.8	95.2	94.5–95.9
RFV or medication (Rx)	303	10	3347	39	88.6	84.8–91.5	99.7	99.5–99.8
RFV or Rx or pathology	342	6	3351	0	100	98.8–100	99.8	99.6–99.9
After duplicates removed	342	6	3341	0	100	98.8–100	99.8	99.6–99.9
Source Practice 3 (n = 7110)^ [T1DM = 26 (21 identified at RFV and 5 at Rx)]								
Reason for Visit (RFV)	242	82	6626	160	60.2	55.3–64.9	98.8	98.5–99.0
RFV or medication (Rx)	385	14	6694	17	95.8	93.3–97.3	99.8	99.6–99.9
RFV or Rx or pathology	402	13	6695	0	100	99.05–100	99.8	99.7–99.9
After duplicates removed	402	13	6644	0	100	99.05–100	99.8	99.7–99.9
Source Practice 4 (n = 12,057)# [T1DM = 47 (35 identified at RFV and 12 at Rx)]								
Reason for Visit (RFV)	590	160	11,065	242	70.9	67.7–73.9	99.6	98.3–98.8
RFV or medication (Rx)	801	26	11,199	31	96.3	94.8–97.4	99.8	99.7–99.8
RFV or Rx or pathology	832	22	11,203	0	100	99.5–100	99.8	99.7–99.9
After duplicates removed	832	22	11,120	0	100	99.5–100	99.8	99.7–99.9
Practice 1 + 2 (n = 4626)) [T1DM = 9 (6 identified at RFV and 3 at Rx)]⁎								
Reason for Visit (RFV)	46	197	4063	320	12.6	9.6–16.4	95.4	94.7–95.9
RFV or medication (Rx)	324	30	4230	42	88.5	84.8–91.4	99.3	99.0–99.5
RFV or Rx or pathology	366	24	4236	0	100	98.9–100	99.4	99.2–99.6
After duplicates removed	302	18	3295	0	100	99.7–100	99.5	99.1–99.7
Practice 1 + 2 + 3 (n = 11,736) [T1DM = 30 (22 identified at RFV and 8 at Rx)]⁎								
Reason for Visit (RFV)	288	279	18,182	480	37.5	34.1–40.9	99.7	99.5–99.8
RFV or medication (Rx)	709	44	10,924	59	92.3	90.2–94.0	99.6	99.5–99.7
RFV or Rx or pathology	768	37	10,931	0	100	99.5–100	99.7	99.5–99.8
After duplicates removed	628	26	8390	0	100	99.4–100	99.7	99.5–99.8
Practice 1 + 2 + 3 + 4 (n = 23,793) [T1DM = 82 (62 identified at RFV and 20 at Rx)]⁎								
Reason for Visit (RFV)	878	439	21,754	722	54.8	52.4–57.3	98.4	98.2–98.6
RFV or medication (Rx)	1510	70	22,123	90	94.4	93.1–95.4	99.6	99.6–99.7
RFV or Rx or pathology	1600	59	22,134	0	100	99.8–100	99.7	99.6–99.8
After duplicates removed	1593	59	21,522	0	100	99.8–100	99.7	99.6–99.8

NOTE: 24 patients had T2DM and T1DM RFV in records (2 from Practice 3 and 22 from Practice 4) and were categorized as T2DM.

Table 2
Accuracy measures by source practice and total data sets (to 3rd level).

Accuracy measures	Practice 1 (*N* = 927)	Practice 2 (*N* = 3699)	Practice 3 (*N* = 7110)	Practice 4 (*N* = 12,057)	Total (*N* = 23,793)
Sensitivity	100.0	100.0	100.0	100.0	100.0
Specificity	98.0	99.8	99.8	99.8	99.7
PPV	57.1	98.3	96.9	97.4	96.4
NPV	98.2	100.0	100.0	100.0	100.0
Accuracy	98.1	99.8	99.8	99.8	99.8
Balanced accuracy	99.0	99.9	99.9	99.9	99.9

4.1. International comparisons

The situation is similar in the authors' countries – Australia, UK and Canada – where similar problems and issues are faced. In the UK the Royal College of General Practitioners (RCGP) and NHS Diabetes set up a classification group to create a pragmatic classification for use in clinical records; the motivation for this was that there were differences between prevalence in epidemiological studies and those found in disease registers, The underlying issue was errors in the coding and classification of diabetes [8,15,35]. This led to an educational initiative and the development of audit tools, made freely available online, customized for each different brand of computer system, to enable practices to improve their data quality (www.clininf.eu/cod/). Superficially, all UK brands of computer system use the same coding system; however, they use different versions, there is variation in the drug dictionary, local codes and how practitioners interface with those codes [36,37]. These are the same problems likely to be encountered in Australia and Canada; though a common data extraction tool usable across all brands of computer system simplifies, though does not solve, the challenge of how to reliably identify people with a specific phenotype [38].

Canada has had similar problems with developing diabetes and other CDM registers [11] despite evidence pointing to the need for one [39]. A foremost issue is that Canada has struggled with EHR design and implementation issues due to overemphasis on data collection and technology rather than meaningful application of EHR data to support essential tasks such as CDM [40].

4.2. Limitations of the method

The manual validation of the positive T2DM cases found in a small general practice may not address the accuracy measure adequately. A PhD project manually audited all the records in the practice and confirmed the accuracy (Sensitivity and Specificity > 95%)

Table 3
Completeness of source practice and multi-practice data set.

Attribute	Completeness* (expressed as% total records)				
	Practice 1 ($N = 927$)	Practice 2 ($N = 3699$)	Practice 3 ($N = 7110$)	Practice 4 ($N = 12,057$)	Total ($N = 23,793$)
All Reason for Visit (RFV)	95.47	86.59	92.14	98.92	94.84
• All DM RFV	4.31	5.65	4.85	6.51	5.8
• T2DM RFV	0.2	1.4	3.7	4.9	3.9
All prescription (Rx)	79.61	94.48	95.82	95.86	95.00
• All diabetes Rx	2.37	8.35	5.37	6.64	6.36
• T2DM Rx	2.6	7.4	3.2	1.8	2.8
All pathology	15.86	61.21	62.95	65.56	62.17
• All HbA1c	0.76	7.98	1.32	1.47	2.41
• HbA1c $\geqslant$ 6.5%	0.54	5.68	0.89	0.98	1.66
• All random BSL	1.29	10.11	45.67	55.04	43.16
• Random BSL $\geqslant$ 11.1	0	0.49	1.39	1.97	1.49
• All fasting BSL	6.15	50.85	8.38	6.3	13.84
• Fasting BSL $\geqslant$ 7	0.32	5.73	0.9	0.75	1.56
All measures	82.31	84.75	90.13	91.95	89.91
• All body mass index	11.54	23.95	19.89	16.12	18.28
• Body mass index $\geqslant$ 30	5.39	10.84	9.68	7.92	8.8
Family History (FH)	18.23	8.57	10.28	9.06	9.7

* Completeness was defined as at least 1 record per patient. For pathology it was based on the test result, prescription on the medication name, and family history and RFV on whether the record contained a text entry.

in this practice; submitted elsewhere as a thesis paper. Finally, data integrity and fidelity issues have not been addressed.

4.3. Next steps

The methodology is valid for quality improvement and research purposes. The integration of biostatistics and informatics methods, including KE, has triangulated and grounded the findings to increase our understanding of the impact of data completeness and dataset size on algorithm accuracy. The ePBRN research program aims to extend this integrative work with data-driven technical and semantic ontological methods. Ontology-based methods are essential to the automated integration and management of the increasing volumes of complex clinical information from disparate EHRs. Basic data quality (DQ) metrics are important and can be embedded in the ontology infrastructure to manage data repositories. We have selected completeness, correctness, consistency and duplication (C3D) as the core DQ metrics for demographic and clinical data collected, and to compare within and among EHRs. Both clinical data and professional knowledge changes over time and ontology-based KE tools are the most promising instruments to deal with this temporal dimension cost-effectively.

Potentially, the concept of ontologically-specified mapping, using -ontopPro- for instance, of relevant data/information may allow automated (intelligent) focused access to distributed EHRs, using only the required data and leaving the other data where they are. This is becoming increasingly necessary when clinical and social media data become so "big" that extracting, moving or copying the data becomes impractical.

Intelligent systems are needed to identify cases of a particular phenotype and integrate patient information to guide practice, research and policy. The ontology-based approach can address the integration of patient data within a single and across multiple EHRs to develop clinical phenotypes for CDM and integration of clinical concepts to model the phenotype (**clinical integration**) recognizing that CDM is non-linear, complex and operates in a context that includes clinical factors such as co-morbidities, risk factors and allergies and non-clinical factors such as health financing and insurance, to support tasks such as clinical care, practice organization, and quality improvement initiatives. Ontological methods also allow a clinician to look at many attributes of a single patient while simultaneously permitting analysts to look at a single attribute across many patients, within the same environment. Models, methodologies and tools to support EHR-driven phenotyping in multiple contexts are needed to support well-coordinated health promotion and prevention and management of chronic disease.

5. Conclusion

We examined the accuracy of a T2DM phenotyping algorithm and its ontology infrastructure in finding cases of T2DM in individual EHRs and a multi-EHR data repository. Our clinical informatics program enables the reuse of the knowledge already represented in SNOMED CT-AU to perform semantic retrievals for different applications and clinical domains. The goal to create knowledge-driven models to integrate disparate datasets and knowledge bases to support the integrated care of patients with multiple chronic diseases and on multiple medications has been partially achieved. The ontologically driven approach can improve the accuracy of EHR-based disease registers and potentially lead to improved and coordinated chronic disease management, patient safety, and quality outcomes.

Funding statement

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. It is part of the electronic Practice Based Research Network (ePBRN) research program which was/is funded in part by UNSW Medicine, School of Public Health & Community Medicine, Ingham Institute for Applied Medical Research and South West Sydney Local Health District.

Competing interests

Apart for STL having intellectual property in the GRHANITE™ tool, there are no known competing interests.

Contributorship statement

STL leads the ePBRN program and this research. HY conducted the infrastructure and technical work. JT conducted most of the clinical research activity for this paper. STL led the writing with conceptual input from SdeL and CZ and biostatistics advice from AH.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2014.07.016.

References

[1] Shaw S, Rosen R, Rumbold B. What is integrated care? Nuffield trust. Report no; June 2011.
[2] Bodenheimer T, Lorig K, Holman H, Grumbach K. Patient self management of chronic disease in primary care. J Am Med Assoc 2002;288(19):2469–75.
[3] Bodenheimer T, Wagner EH, Grumbach K. Improving primary care for patients with chronic illness. J Am Med Assoc 2002;288(15):1909–14.
[4] Hibbard J, Mahoney E, Stockard J, Tusler M. Development and testing of a short form of the patient activation measure. HSR: Health Serv Res 2005;40(6 (Part 1)):1918–30.
[5] Chute C, Pathak J, Savova G, Bailey K, Schor M, Hart L, et al., editors. The SHARPn project on secondary use of electronic medical record data: progress, plans, and possibilities. AMIA annual symposium 2011. Washington (DC); 2011.
[6] Coiera E. Social networks, social media, and social diseases. BMJ 2013;346(f3007) [22 May, 2013].
[7] Wei W, Leibson C, Ransom J, Kho A, Caraballo P, Chai H, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. J Am Med Inform Assoc 2012;19(2):219–24 [Pubmed Central PMCID: PMC3277630. Epub 2012 January 16, 2012 March–April].
[8] de Lusignan S, Sadek N, Mulnier H, Tahir A, Russell-Jones D, Khunti K. Miscoding, misclassification and misdiagnosis of diabetes in primary care. Diabet Med 2012;29(2):181–9.
[9] Martin D, Wright J. Disease prevalence in the English population: a comparison of primary care registers and prevalence models. Soc Sci Med 2009;68(2):266–74.
[10] Ford D, Knight A. The Australian primary care collaboratives: an Australian general practice success story. Med J Aust 2010;193(2):90–1 [July 19, 2010].
[11] Ferguson R. Auditor general's report: scrapped diabetes registry cost Ontario government $24.4 million. The Star 2013;2012 [April 08, 2013].
[12] Langan S, Benchimol E, Guttmann A, Moher D, Petersen I, Smeeth L, et al. Setting the RECORD straight: developing a guideline for the REporting of studies conducted using observational routinely collected data. Clin Epidemiol 2013;5:29–31.
[13] Liyanage H, Liaw ST, de Lusignan S. Reporting of studies conducted using observational routinely collected data (RECORD) statement: call for contributions from the clinical informatics community. Inform Prim Care 2012;20(4):221–4.
[14] Mehta A. The how (and why) of disease registers. Early Hum Develop 2010;86(11):723–8.
[15] de Lusignan S, Khunti K, Belsey J, Hattersley A, van Vlymen J, Gallagher H, et al. A method of identifying and correcting miscoding, misclassification and misdiagnosis in diabetes: a pilot and validation study of routinely collected data. Diabet Med 2010;27:203–9.
[16] Liaw S, Taggart J, Dennis S, Yeo A. Data quality and fitness for purpose of routinely collected data – a case study from an electronic practice-based research network (ePBRN). In: American medical informatics association annual symposium 2011. Washington (DC): Springer Verlag; 2011.
[17] Taggart J, Liaw S, Dennis S, Yu H, Rahimi A, Jalaludin B, et al., editors. The University of NSW electronic practice based research network: disease registers, data quality and utility. In: 20th Australian national health informatics conference (HIC 2012). Sydney: Studies in Health Technology and Informatics; 2012.
[18] Gruber TR. Toward principles for the design of ontologies used for knowledge sharing. Int J Hum Comput Stud 1995;43(5–6) [Epub 928].
[19] Carlsson S, editor. A critical realist perspective on IS evaluation research. European conf information systems 2005 proceedings; 2005.
[20] Pawson R, Tilley N. Realistic evaluation. London: Sage; 1997.
[21] Feigenbaum EA, McCorduck Pamela. The fifth generation. 1st ed. Reading (MA): Addison-Wesle; 1983.
[22] Colombo G, Merico D, Boncoraglio G, De Paoli F, Ellul J, Frisoni G, et al. An ontological modeling approach to cerebrovascular disease studies: the NEUROWEB case. J Biomed Inform 2010;43(4):469–84.
[23] Beck T, Gollapudi S, Brunak S, Graf N, Lemke H, Dash D, et al. Knowledge engineering for health: a new discipline required to bridge the "ICT gap" between research and healthcare. Hum Mutat 2012;33(5):797–802.
[24] Liaw S, Boyle DIR. Primary care informatics and integrated care of chronic disease. In: Hovenga EKM, Garde S, Cossio CHL, editors. Health informatics: an overview. Studies in health technology and informatics, vol. 151. IOSPress; 2010.
[25] Liyanage H, Liaw S, Kuziemsky C, de Lusignan S, editors. Ontologies to improve chronic disease management research and quality improvement studies – a conceptual framework. Medinfo 2013. Copenhagen; 2013.
[26] Colagiuri S, Davies D, Girgis S, Colagiuri R. National evidence based guideline for case detection and diagnosis of type 2 diabetes. Canberra: Diabetes Australia and the NHMRC; 2009.
[27] Dunstan D, Zimmet P, Welborn T, De Courten M, Cameron A, Sicree R, et al. The rising prevalence of diabetes and impaired glucose tolerance: the Australian diabetes, obesity and lifestyle study. Diabetes Care 2002;25(5):829–34.
[28] Rubin D, Lewis S, Mungall C, Misra S, Westerfield M, Ashburner M, et al. National center for biomedical ontology: advancing biomedicine through structured organization of scientific knowledge. OMICS (Summer) 2006;10(2):185–98.
[29] Perez-Rey D, Maojo V, Garcia-Remesal M, Alonso-Calvo R, Billhardt H, Martin-Sanchez F, et al. ONTOFUSION: ontology-based integration of genomic and clinical databases. Comput Biol Med 2006;36(7–8):712–30 [PubMed PMID: 16144697. Epub 2005/09/08. Eng].
[30] de Lusignan S, Liaw S, Michalakidis G, Jones S. Defining datasets and creating data dictionaries for quality improvement and research in chronic disease using routinely collected data: an ontology-driven approach. Inform Prim Care 2011;19(3):127–34.
[31] Rodriguez-Muro M, Calvanese D. Quest, a system for ontology based data access. KRDB research centre for knowledge and data: free University of Bozen-Bolzano; 2012.
[32] Kahn M, Raebel M, Glanz J, Reidlinger K, Steiner J. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care 2012;50(7 Suppl 1). S2–S29.
[33] Newcombe R. Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med 1998;17:857–72.
[34] Daneman D. Type 1 diabetes. Lancet 2006;367(9513):847–58.
[35] Sadek K, Khunti K, de Lusignan S. Classification of diabetes for primary care: a practical approach. Diabetes Prim Care 2012;14(5):284–92.
[36] Tai T, Anandarajah S, Dhoul N, de Lusignan S. Variation in clinical coding lists in UK general practice: a barrier to consistent data entry? Inform Prim Care 2007;15(3):143–50.
[37] Michalakidis G, Kumarapeli P, Ring A, van Vlymen J, Krause P, de Lusignan S. A system for solution-orientated reporting of errors associated with the extraction of routinely collected clinical data for research and quality improvement. Stud Health Technol Inform 2010;160(Pt 1):724–8.
[38] de Lusignan S, Valentin T, Chan T, Hague N, Wood O, van Vlymen J, et al. Problems with primary care data quality: osteoporosis as an exemplar. Inform Prim Care 2004;12(3):147–56.
[39] Lipscombe L, Hux J. Trends in diabetes prevalence: incidence and mortality in Ontario, Canada 1995–2005: a population-based study. Lancet 2007;369:750–6.
[40] Rozenblum R, Jang Y, Zimlichman E, Salzberg C, Tamblyn M, Buckeridge D, et al. A qualitative study of Canada's experience with the implementation of electronic health information technology. CMAJ 2011;183(5):E281–8.