Siaw-Teng Liaw
Jane Taggart
Hairong Yu
Simon de Lusignan

Data extraction from electronic health records – existing tools may be unreliable and potentially unsafe

Background

The increasing use of routinely collected data in electronic health record (EHR) systems for business analytics, quality improvement and research requires an extraction process fit for purpose. Little is known about the quality of EHR data extracts. We examined the accuracy of three data extraction tools (DETs) with two EHR systems in Australia.

Methods

The hardware, software environment and extraction instructions were kept the same for the extraction of relevant demographic and clinical data for all active patients with diabetes. The counts of identified patients and their demographic and clinical information were compared by EHR and DET.

Results

The DETs identified different numbers of diabetics and measures of quality of care under the same conditions.

Discussion

Current DETs are not reliable and potentially unsafe. Proprietary EHRs and DETs must support transparency and independent testing with standardised queries. Quality control within an appropriate policy and legislative environment is essential.

Keywords

electronic health record; quality of health care; medical informatics

Current health reforms promote electronic health records (EHRs)[1–3] to monitor the quality and safety of care[4] and research.[5] Practice-based clinical datasets are increasingly being extracted into data repositories to be mined for business analytics,[6] research[7] and quality improvement,[8] making it possible to measure quality and health outcomes on a scale and at a speed not possible with manual records. However, such data analytics are limited by the quality of the data recorded, and EHRs may impose their own limitations.[9] While data have been extracted from EHRs for two decades, we know little about the quality of the EHR data extracts or accuracy of the data extraction tools (DET) used.

Commercial DETs exist, but, like EHRs, they are largely proprietary 'black-box' solutions with intellectual property protection preventing adequate assessment of any design or execution errors or quality of data extracted. Effective assessment and management of data quality (DQ) requires analysis of the whole data cycle: from collection through extraction, cleansing, storage, management, dissemination, presentation and curation.[10] Data quality management (DQM) processes and information governance (IG) structures are needed to ensure that data routinely captured within clinical practice is complete, correct, consistent[11] and, ultimately, is fit for purpose.[12]

We examined whether different DETs achieved consistent results. Diabetes was used as the exemplar because it has a known prevalence, is clinically important and should be consistently extracted from EHRs as the diagnosis is based on numeric data[13] and most anti-diabetic drugs and pathology tests are diabetes-specific. United Kingdom researchers have set out the sensitivity and specificity of surrogate markers of diabetes[14] and differentiated between people with poor DQ within their EHRs, subdividing them into those who have errors in coding, classification or diagnosis of their diabetes. Around 40% of people with one or more of these errors have underlying clinically significant issues[15] and those not included in computerised patient registers seem to receive worse care.[16]

The University of New South Wales (UNSW) electronic Practice-Based Research Network (ePBRN)[7] compared 'DET1', its in-house data extraction and linkage tool,[17,18] to two other DETs. We tested the hypothesis that the counts of the diabetes cases identified and their demographic and clinical data extracted from a general practice EHR will be the same for DET1 and two other proprietary DETs.

Methods

Two different EHRs (EHR1 and EHR2), were used to compare DET1, which extracts from both EHRs, with:

1. DET2, which extracts from EHR1 and EHR2
2. DET3, which only extracts from EHR1.

EHR1 uses a proprietary coding (terminology) system while EHR2 uses the International Classification of Primary Care version 2 (ICPC2).[19] Both EHRs allow free text entry if the codes are not available. Electronic health records are typically relational databases with a number of linked data tables, including: 'History', 'Past History', 'Diagnosis', 'Medication', 'Pathology', 'Measures' and so on.

The data models vary as there are no prescribed standards. In EHR1, the 'Diagnosis Table' and 'History Table' capture diagnostic terms from 'reason for visit', 'prescription' and 'procedure' data entry windows.

The location of key-coded data varies from EHR to EHR, and clinicians may be unclear as to the consequences of their recording choices, or even of what the default settings might do. The following examples set out how apparently innocent choices about where data are stored or saved might affect whether a case, risk factor or treatment might be identified by a subsequent search:

- The 'reason for contact' that can include a problem or diagnosis is saved in the 'Diagnosis Table'. If 'save in past history' is ticked, the default option in EHR1, it is also saved in the 'History Table'. In EHR2, the ICPC2 codes are captured in the 'Problem Table' and glycosylated haemoglobin (HbA1c) tests in the 'Pathology Table'.
- DETs often use different tables or a number of tables as the data source, eg. DET2 and DET3 identified diabetes cases from the 'condition' attribute in the 'History Table' of EHR1, while DET1 used the 'reason for contact' attribute in the 'Diagnosis Table'. DET2 and DET3 also 'transformed' the extracted data, using a proprietary tool.

The extractions were run one after the other for each EHR (*Figure 1*). DET1 used an extensible mark-up language (XML)-specified query to extract data. For DET2, relevant terms were entered into the search window to identify patients flagged as active in the EHR (EHR-active) and pre-programmed reports generated. For DET3 the pre-programmed reports were executed. All extracted data were sent to a data repository where they were analysed using Statistics Package for the Social Sciences (SPSS) software (version 20.0; IBM SPSS, Armonk, NY, USA). DET1 used a comprehensive list of approximately 300 structured and free text terms used to describe diabetes in both EHRs; DET2 and DET3 used the relevant coded terms available for EHR1, but the documentation was not explicit about the actual terms used. For EHR2, DET1 and DET2 used ICPC2[19] codes for diabetes and diabetes-related conditions (T89, T90, T31, T45, T68, T99, N94, F83 and the relevant 6-digit extensions).

The causes of the variations were categorised into clinical practice, EHR, DET and data quality factors. This built on an existing taxonomy of information technology (IT)-related data extraction errors,[20] which categorised errors in extracted data into: (1) extraction method and process, (2) translation layer, (3) shape and complexity of original schema, (4) communication and system (software) faults, (5) hardware and infrastructure, and (6) generic or human errors.

Results

The three DETs delivered different counts of 'EHR-active' patients, diabetics and quality (completeness) of diabetes data (*Table 1*). Differences ranged from 0.1% to 10%, some statistically significant, for active patients; diabetics identified by condition; HbA1c; diabetes medication; and risk factors such as smoking status, body mass index (BMI) and blood pressure (BP).

EHR-active patients

The number of designations for EHR-active patients differed significantly, with up to 10% difference between the DETs. Possible explanations include incomplete or untimely flagging of EHR-active status by staff/general practitioners (GPs); inconsistent use of patient status codes (such as death, transferred or deleted) and search queries (such as how the DET/EHR handles the options for patient status, empty data fields or the 'next-of-kin' field); incompatibility of the DET/EHR data models; corrupted interactions of the DET/EHR software routines; and data corruption for technical and functional reasons.

Identifying diabetes by specific database tables

The number of diabetics identified by a diagnostic label varied by DET and source table used. (a) EHR1: DET1 identified more diabetics than the other DETs from the 'History Table' but fewer from the 'Diagnosis Table'. Because DET2 and DET3 used the same coded terms, they extracted nearly identical numbers from the 'History Table'. DET1 extracted more medications than DET3 from EHR1, largely due to differences in terms/codes used for the diabetes medication queries. DET1 used the therapeutic class (THERCLASS) codes of EHR1, whereas DET3 identified medications in EHR1 by what they called 'mappings' to the THERCLASS but did not provide details. (b) EHR2: DET1 extracted more diabetics (using ICPC2 codes) than DET2 from the 'Problem Table' and HbA1c tests from the 'Pathology Table'.

Identifying diabetes using multiple tables

By combining the information from multiple tables ('Diagnosis', 'Pathology', 'Prescriptions' and

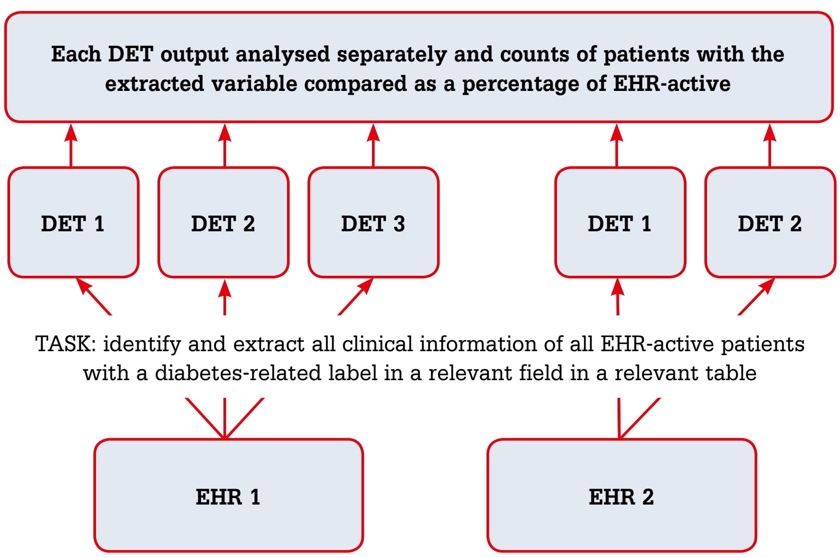

Figure 1. Methodology for comparison of DETs output

'Measures') and eliminating duplicates, DET1 identified more diabetics than when using a single table (*Table 1*). DET2 and DET3 identified patients only from the 'History' table.

Demographics

A similar random variation pattern was found, which may reflect differences in the diabetics identified, data-entry options for gender/date of birth or how the EHR handles data storage and exchange between EHRs and linked billing systems, usually separate proprietary software. Incomplete or incorrect data entry may be an issue as one patient had an EHR default date of '01 Jan 1800' in the date-of-birth field.

Risk factors

All three DETs extracted different numbers of diabetics with a recorded BMI or BP from EHR1. When expressed as a proportion of the numbers of diabetics the DET identified, DET1 extracted a lower proportion across all the risk factors.

Discussion

The DET/EHR data models were proprietary and not transparent. However, we gained sufficient insight into the DET/EHR from available documentation, iterative use and discussions with vendor technical support and members of the ePBRN. Generally, DETs extract, store, manipulate and report on a snapshot of coded data in the EHR. The data model, data and metadata are not tailored or updated systematically or validated independently or in association with the EHR. As such, there are often mismatches between the DET and EHR data models. The ePBRN experience with repeated data extractions suggests that EHRs are often not consistent in how they store codes or data over time.

The sociotechnical conceptual framework describes how users and technology undergo a process of mutual transformation. Flagging variations in extracted data in terms of technical and system design factors, differing practices in documentation, workflow and related factors may improve system design and more consistent recording of data.[21] This study highlights the variation between each DET/EHR combination, and sets a potential agenda for improving our ability to monitor quality.

A limitation of this study is that it reported crude extract numbers. None provided a 'gold standard' extraction that matches the expected 6.6% prevalence of diabetes in the study regions, as reported in the South Western Sydney Local Health District 2012 Annual Report. While we do not provide adjusted prevalence figures, it is plausible that some DETs are missing large numbers of relevant patients.

We conclude that the DET/EHR combinations did not extract similar counts of diabetics and indicators of diabetes care. This renders current DETs ineffective as tools for measuring the quality of care in a way that might be compared between systems. When we add the lack of transparency for proprietary reasons and a lack of technical and professional standards and safety regulations for medical software, this situation is unable to ensure that practice is safe, or able to support clinical governance.

Organisations promoting eHealth must be accountable and transparent[22] and their software products subject to appropriate and independent accreditation and regular review, including monitoring for critical incidents associated with their use.

Implications for general practice

- Data extracted from EHRs may be unreliable.
- EHRs and extraction tools must support independent testing with standardised queries.
- The proprietary model for software quality control is not in our best interests.
- Appropriate information governance processes and structures must be established.

Table 1. Variations in a sample of data extracted from two EHRs by three DETs under the same experimental conditions

Data extracted from specific tables in EHR:	DET1		DET2		DET3	
	Patients n (%)	95% Confidence intervals (CI)	Patients n (%)	95% CI	Patients n (%)	95% CI
1. EHR1						
All EHR-active patients	21 793	N.A.	24,145	N.A.	24,180	N.A.
Patients with diabetes-related diagnostic labels	558 (2.9)	2.69–3.13	598 (2.5)	2.29–2.68	599 (2.5)	2.29–2.68
Patients with HbA1c tests	296 (1.4)	1.21–1.52	253 (1.0)	0.90–1.80	253 (1.0)	0.93–1.18
Patients with diabetes-related prescriptions	642 (3.0)	2.73–3.18	Did not extract routinely		485 (2.0)	1.84–2.19
Patients with diabetes-related diagnostic labels, tests or prescriptions	833 (3.8)	3.60–4.10	DET2 and DET3 do not routinely extract from multiple tables			
2. EHR2						
All EHR-active patients	25 770	N.A.	25,770	N.A.	DET3 does not extract from EHR2	
Patients with diabetes-related diagnostic labels	367 (1.4)	1.29–1.58	234 (0.9)	0.80–1.03		

Authors

Siaw-Teng Liaw PhD, FRACGP, FACHI, is Professor of General Practice, University of New South Wales, and Director, General Practice Unit, South West Sydney Local Health District, NSW. siaw@unsw.edu.au

Jane Taggart MPH, BEd, Dip PE, is a Research Fellow, Centre for PHC and Equity, University of New South Wales, Sydney, NSW

Hairong Yu BEng (Hons), MSc, PhD, is a Senior Research Fellow, Centre for PHC and Equity, University of New South Wales, Sydney, NSW

Simon de Lusignan MBBS, MSc, MD (Res), FHEA, FBCS, CITP, FRCGP, is Professor of Primary Care and Clinical Informatics, and Head, Department of Health Care Management and Policy, The University of Surrey, Guildford, UK

Competing interests: Siaw-Teng Liaw has intellectual property in DET1, but was not involved in the conduct of the experiment reported in this study. This research was funded in part by a seeding grant from the School of Public Health & Community Medicine, University of New South Wales.

Ethics approval: This study has approval from the UNSW Human Research Ethics Committee.

Provenance and peer review: Not commissioned; externally peer reviewed.

Acknowledgements

The authors wish to thank Dr Douglas Boyle, Mr Ian Peters and Ms Yin Huynh for advice on the DETs, and Dr Blanca Gallego-Luxan and Prof Simon Jones for comments on drafts.

References

1. National Health & Hospital Reform Commission. A Healthier Future For All Australians – Final Report of the National Health and Hospitals Reform Commission – June 2009. Canberra: Commonwealth of Australia 2009. Report No. P3 -5499.
2. Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. J Am Med Inform Assoc 2007;14:1–9.
3. Blumenthal D, Tavenner M. The 'meaningful use' regulation for electronic health records. N Engl J Med 2010;363:501–04.
4. Pérez-Cuevas R, Doubova S, Suarez-Ortega M, et al. Evaluating quality of care for patients with type 2 diabetes using electronic health record information in Mexico. BMC Med Inform Decis Mak 2012;12:50.
5. Economou A, Grey M, McGregor J, et al. The health informatics cohort enhancement project (HICE): using routinely collected primary care data to identify people with a lifetime diagnosis of psychotic disorder. BMC Res Notes 2012;5:95.
6. Adams J, Klein J. Business Intelligence and Analytics in Health Care - A Primer 2011: Available at www.advisory.com/Research/IT-Strategy-Council/Research-Notes/2011/Business-Intelligence-and-Analytics-in-Health-Care [Accessed 7 August 2013]
7. Taggart J, Liaw S, Dennis S, et al. The University of NSW electronic Practice Based Research Network: Disease registers, data quality and utility. 20th Australian National Health Informatics Conference (HIC 2012). Sydney: Studies in Health Technology and Informatics; 2012.
8. Ford D, Knight A. The Australian Primary Care Collaboratives: an Australian general practice success story. Med J Aust 2010;193:90–91.
9. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013;20:117–21.
10. de Lusignan S, Liaw S, Krause P, et al. Key concepts to assess the readiness of data for International research: Data quality, lineage and provenance, extraction and processing errors, traceability, and curation. Yearb Med Inform 2011;6:112–21.
11. 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). American Medical Informatics Association Annual Symposium 2011; Washington DC: Springer Verlag; 2011.
12. Liaw S, Chen H, Maneze D, et al. Health reform: is routinely collected electronic information fit for purpose? Emerg Med Australas 2012;24:57–63.
13. World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: report of a WHO/IDF consultation. Geneva: World Health Organization; 2006.
14. Bagheri A, Sadek A, Chan T, Khunti K, de Lusignan S. Using surrogate markers in primary electronic patient record systems to confirm or refute the diagnosis of diabetes. Inform Prim Care 2009;17:121–29.
15. 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;29:181–89.
16. Hassan Sadek N, Sadek A, Tahir A, Khunti K, Desombre T, de Lusignan S. Evaluating tools to support a new practical classification of diabetes: excellent control may represent misdiagnosis and omission from disease registers is associated with worse control. Int J Clin Pract 2012;66:874–82.
17. Liaw S, Boyle DIR. Primary care informatics and integrated care of chronic disease. In: Hovenga E KM, Garde S, Cossio CHL, editors. Health informatics: An overview. Amsterdam: IOS Press; 2010.
18. Boyle D, Rafael N. BioGrid Australia and GRHANITE™: privacy-protecting subject matching. Stud Health Technol Inform 2011;168:24–34.
19. Classification Committee W. ICPC-2-R: International Classification of Primary Care. Revised, 2nd edn. Oxford: Oxford University Press; 2005.
20. Michalakidis G, Kumarapeli P, Ring A, van Vlymen J, Krause P, de Lusignan S, editors. A system for solution-orientated reporting of errors associated with the extraction of routinely collected clinical data for research and quality improvement. MEDINFO 2010; Cape Town, South Africa: IOS Press; 2010.
21. Parsons A, McCullough C, Wang J, Shih S. Validity of electronic health record-derived quality measurement for performance monitoring. J Am Med Inform Assoc 2012;19:604–09.
22. Fisher ES, McClellan MB, Safran DG. Building the path to accountable care. N Engl J Med 2011;365:2445–47.

correspondence **afp@racgp.org.au**