International Journal of Population Data Science

Journal Website: www.ijpds.org

Understanding data provenance when using electronic medical records for research: Lessons learned from the Deliver Primary Healthcare Information (DELPHI) database

Jason Edward Black[1], Amanda L. Terry[1,2,3], Sonny Cejic[1], Tom Freeman[1], Dan Lizotte[2,3,4], Scott McKay[1], Mark Speechley[2,3], and Bridget Ryan[1,2,3]

Submission History

Submitted:	**21/06/2023**
Accepted:	**30/08/2023**
Published:	**28/09/2023**

[1]Department of Family Medicine, Schulich School of Medicine and Dentistry, Western University, London, ON, Canada
[2]Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, Western University, London, ON, Canada
[3]Schulich Interfaculty Program in Public Health, Schulich School of Medicine and Dentistry, Western University, London, ON, Canada
[4]Department of Computer Science, Faculty of Science, Western University, London, ON, Canada

Abstract

Introduction
We set out to assess the impact of Choosing Wisely Canada recommendations (2014) on reducing unnecessary health investigations and interventions in primary care across Southwestern Ontario.

Methods
We used the Deliver Primary Healthcare Information (DELPHI) database, which stores deidentified electronic medical records (EMR) of nearly 65,000 primary care patients across Southwestern Ontario. When conducting research using EMR data, data provenance (i.e., how the data came to be) should first be established. We first considered DELPHI data provenance in relation to longitudinal analyses, flagging a change in EMR software that occurred during 2012 and 2013. We attempted to link records between EMR databases produced by different software using probabilistic linkage and inspected 10 years of data in the DELPHI database (2009 to 2019) for data quality issues, including comparability over time.

Results
We encountered several issues resulting from this change in EMR software. These included limited linkage of records between software without a common identifier; data migration issues that distorted procedure dates; and unusual changes in laboratory test and medication prescription volumes.

Conclusion
This study reinforces the necessity of assessing data provenance and quality for new research projects. By understanding data provenance, we can anticipate related data quality issues such as changes in EMR data over time—which represent a growing concern as longitudinal data analyses increase in feasibility and popularity.

Keywords
electronic medical records; DELPHI; Choosing Wisely Canada; longitudinal analysis; interrupted time series; data quality; data provenance; data linkage

*Corresponding Author:
Email Address: jblack85@uwo.ca (Jason Edward Black)

https://doi.org/10.23889/ijpds.v8i5.2177

Key features

- In this Data Note, we demonstrate the importance of inspecting data provenance (i.e., how the data came to be) when considering electronic medical records (EMR) for research use.
- We set out to assess the impact of Choosing Wisely Canada recommendations (2014) on reducing unnecessary health interventions in primary care across Southwestern Ontario.
- We used the Deliver Primary Healthcare Information (DELPHI) database, which stores deidentified EMRs of nearly 65,000 primary care patients across Southwestern Ontario.
- Records were stored by one EMR software up to 2012 and a different software thereafter. We attempted to link records between EMR databases produced by different software with limited success.
- We encountered several issues resulting from this change in EMR software. These included limited linkage of records between software without a common identifier; data migration issues that distorted procedure dates; and unusual changes in laboratory test and medication prescription volumes.
- By understanding data provenance, we can anticipate related data quality issues such as changes in EMR data over time—which represent a growing concern as longitudinal data analyses increase in feasibility and popularity.

Introduction

Electronic medical records (EMRs) have emerged as a prominent data source for health research [1, 2]. While collected for administration and patient care purposes, data from primary care EMRs can be used to gain powerful insights about the health and healthcare of primary care populations [3, 4]. For each patient visit, detailed information about the visit—including date, time, patient clinical characteristics, laboratory test results, and procedures performed—is recorded as a unique entry in the EMR, allowing for longitudinal analysis to observe changes in patient health and primary care services over time [5], such as patterns in the accumulation of chronic conditions (i.e., multimorbidity) [6].

In primary care research, patient records that are input by practitioners and stored by the practice site EMR software are periodically extracted and stored by regional networks across Canada using strict security and privacy protocols. Extensive tools and strategies for analysing EMR data have been developed to support EMR-based research, such as disease case definitions [7]. Despite these tools, EMR data are not collected primarily for research purposes; therefore, it is essential to determine the suitability of EMR data for research prior to analysis by assessing data provenance (i.e., how the data came to be) and data quality. While methods for assessing EMR data quality are available [8–12], data provenance is frequently overlooked.

One crucial component of data provenance is the software used to store patient records and any software changes over time. Many EMR software exist to store patient records in Canada, including several private and open-source options (e.g., Accuro EMR, Med Access, and PS Suite EMR). The software used to store primary care data determines several aspects of data provenance that impact analysis, including what data can be stored (e.g., diagnostic billing codes, risk factor information, medication prescriptions); what format the data are stored (e.g., structured responses or unstructured text); units of measurement (e.g., pounds or kilograms); and sometimes what values can be stored (i.e., validity checks may prevent entering implausible values). For example, if the software does not offer a structured medication prescription field, this must be considered when conducting analyses focused on prescribing across practices using different software. Further, primary care practice sites may change EMR software over time, which can impact the patient information that is stored. For example, introducing a newer version of diagnostic coding (e.g., ICD-10-CA replacing ICD-9) will create differences in the level of detail in diagnostic data. Additionally, changing software requires migration of existing patient records to the new software and/or linkage of existing patient records with records generated by the new software.

In this article, we describe how aspects of data provenance, such as the migration from one EMR software to another, can impact EMR data that are extracted for research and the implications for analysis—with a focus on longitudinal analyses. We present as a case study an analysis of data from the Deliver Primary Healthcare Information (DELPHI)[1] EMR database in Southwestern Ontario, Canada [13].

Background

The national Choosing Wisely Canada (CWC)[2] initiative was launched on April 2nd, 2014 with the aim to reduce unnecessary tests, treatments, and procedures in Canadian health care [14, 15]. We set out to assess the impact of several CWC recommendations (Table 1) on reducing unnecessary health interventions in primary care across Southwestern Ontario.

For this study, we analysed DELPHI primary care EMR data spanning January 2009 to March 2019. We planned an *interrupted time series analysis* to observe changes in patterns of *indicators* that could be attributed to an *interruption,* especially population-level interventions. We evaluated the impact of all CWC recommendations released in April 2014 (Table 1), excluding their recommendation on Pap smears, due to a change in how practitioners billed for Pap smears around this time in Ontario that prevented us from attributing changes in Pap smear rates to CWC recommendations. We compared the 5-year periods before and after the CWC recommendations were released in April 2014.

[1] https://www.schulich.uwo.ca/familymedicine/research/csfm//research/current_projects/delphi.html.

[2] https://choosingwiselycanada.org/.

Table 1: Choosing Wisely Canada recommendations investigated

Recommendation	Indicator
Don't do imaging for lower-back pain without serious underlying conditions	X-rays for lower-back pain
Don't use antibiotics for upper respiratory infections that are likely viral in origin	Antibiotics for upper respiratory tract infections
Don't do chest x-rays and electrocardiographs (ECGs) for asymptomatic or low-risk outpatients	Chest x-rays and electrocardiographs
Don't do annual screening blood tests unless directly indicated by the risk profile of the patient	Screening blood tests

Methods

Launched in 2004, the DELPHI database stores deidentified records of nearly 65,000 primary care patients contributed by 60 practitioners among 18 practice sites across Southwestern Ontario for research use [13]. The DELPHI database comprises consenting patient records including diagnoses, procedures, laboratory tests, medication prescriptions, and referrals as recorded by participating primary care practitioners in their EMR. Records are extracted on a regular basis (approximately semi-annually) from participating physicians' EMR databases and processed for research use, including deidentification, cleaning and coding (e.g., converting free-text fields into corresponding diagnostic codes), and standardisation to a common format. Research using DELPHI data has included studies on musculoskeletal conditions, congestive heart failure, and patterns of referral [16–18]; each study examined the quality of the data prior to conducting the analyses. The time period for these studies typically spanned no more than five years.

We considered DELPHI data provenance and quality, and how it may impact our proposed longitudinal analyses. Knowing a change in EMR software occurred during 2012 and 2013 among all DELPHI practices, we inspected 10 years of data. To begin, we attempted to link patient records between EMR databases before and after the software change. No common unique identifier existed between software; as such, we conducted probabilistic matching based on patient characteristics. Subsequently, we quantified the incidence of each indicator before and after the change in software. To identify occurrences of each indicator in our EMR data source, we developed case definitions based on combinations of diagnostic codes, free-text descriptions, and medication names in collaboration with our physician co-investigators who have used EMRs in clinical practice (see Appendix).

Results

As recommended by Reimer et al. [12], we first determined the proportion of patient records stored by the previous EMR software that could be linked with records stored by the current EMR software within each practice. No unique identifier existed that spanned both databases produced by the previous and current EMR software. As such, we relied on probabilistic matching, where multiple data fields (i.e., age, gender, geographic region, and primary care encounter date) attempted to uniquely identify patients within both EMR databases produced by different software. Based on probabilistic matching, we were able to match 36.4% of patients within the database produced by current EMR software to their records stored within the database produced by the previous EMR software. Matched patients were similar to unmatched in terms of age and sex/gender (Table 2); however, matched patients more frequently lived in rural settings, indicating strong geographic differences. Though this low proportion of patients that were successfully linked prohibited us from following most individual patients over time, we could still examine the frequency of our indicators over time.

Next, we considered the longitudinal concordance between databases produced by different software for our indicators (i.e., "data element presence, agreement, and source agreement of specified variables across multiple data sources" [12]). For example, when examining the records of medical procedures ordered or conducted by practitioners (e.g., diagnostic imaging or electrocardiography), the years 2012 and 2013 contained 2-3 times more procedures recorded than subsequent years—and no procedures were recorded before 2012 (Table 3). In the process of migrating data between EMR databases produced by different software in 2012 and 2013, we inferred that procedure dates may have been inappropriately assigned the date of migration rather than the date the procedure was conducted. This data quality issue precluded our analysis of two procedure indicators (i.e., x-rays for lower-back pain and chest x-rays and electrocardiographs). Similarly, medication prescription data volume varied before and after data migration precluding its analysis. For example, far fewer antibiotic prescriptions per year were identified after the change in EMR software (475 prescriptions per year after vs. 1,601 prescriptions per year before the change) despite containing approximately the same number of patient encounters per year. Accordingly, we were not confident that the newer EMR software captured all medication prescriptions.

Prompted by our knowledge of DELPHI data provenance, we identified data quality issues resulting from a change in EMR software, such as limited linkage of patient records between databases produced by the current and prior EMR software and migration issues that distorted procedure dates. This inspection of data provenance and quality determined that the DELPHI database was not suitable to assess the impact of the CWC recommendations.

Table 2: Characteristics of patients from the database produced by the new EMR software matched and unmatched with records from the database produced by the previous EMR software

Characteristic	Matched patients (n=5,807)	Unmatched patients (n=10,166)
Age, mean (SD)	39.1 (21.4)	39.2 (21.2)
Sex/gender, n (%)		
Male	3,057 (52.6)	5,390 (53.0)
Female	2,750 (47.4)	4,773 (47.0)
Unknown	0	3 (0.03)
Rural/urban, n (%)		
Rural	2,799 (48.2)	3,040 (29.9)
Urban	1,997 (34.4)	5,576 (54.9)
Unknown	1,011 (17.4)	1,550 (15.3)

Table 3: Frequency of procedures in the database produced by the newer EMR software *Procedures* table, by year

Year	Frequency (n)
2009	0
2010	0
2011	0
2012	17,171
2013	12,867
2014	5,184
2015	4,244
2016	5,724
2017	4,207
2018	3,361
2019	761

Discussion

This paper reports our examination of data provenance in the DELPHI EMR database for a longitudinal analysis. Our findings underline the importance of establishing an essential aspect of EMR data provenance among many considerations when assessing a population database and its metadata for research use [19, 20]: adoption of a new EMR software and migration of data from the previous software. It is a natural expectation that practices may change their choice of EMR software vendors to address changing needs, or as some vendors disappear from the marketplace. Even smaller version updates to existing EMR software may appreciably impact how data are collected and stored. This alters data provenance and has its largest impact when conducting longitudinal studies.

Understanding data provenance helps to encourage improved data quality; audit trails; reproducibility of data processing; proper attribution and ownership of the data; and discovery of new information about the processing [20]. Changes in EMR software may impact the quality of the data primarily but also must be understood to allow reproducibility. An understanding of data provenance obtained by reviewing relevant metadata and consulting EMR users and data managers to help identify areas of concern for further data quality investigation. Face validity of study results (i.e., whether results are reasonable based on existing knowledge) may not sufficiently identify all data quality issues. In our case, working with the DELPHI data manager helped identify a change in software over 2012 and 2013 and allowed us to anticipate and identify data quality issues that precluded our ability to conduct a proposed study of the impact of the CWC recommendations.

Inspecting data provenance can be facilitated by reviewing metadata provided for the database. Indeed, proper metadata documentation is crucial, though commonly insufficient or absent [19]. Importantly, we advise researchers to inspect available metadata and/or work with EMR data managers to identify any EMR software changes as part of their data provenance considerations when conducting data quality evaluation and analysis. Specifically, when a change in EMR software occurred, we recommend understanding 1) all relevant transition dates, including the date of change in software and data migration dates; 2) which data were migrated and stored by the new software and which were not; and 3) what information uniquely identifies each patient in data stored by both software—ideally this is an assigned unique identifier (e.g., *Patient ID*) but can sometimes be approximated by a unique combination of patient characteristics (e.g., date of birth, geographic region, gender, etc.). With this information, EMR data can be inspected for data quality issues resulting from changes in software and potentially remediated to allow for valid inferences. Our experiences highlight the need for further research aimed at establishing a broad set of data provenance criteria (including but not limited to changes in EMR software) to be evaluated prior to conducting research using EMR data, similar to those proposed for EMR data quality [8–12].

Conclusions

By understanding data provenance, we can anticipate related data quality issues such as changes in EMR data over time—which represent a growing concern as longitudinal data analyses increase in popularity and EMR databases age to contain multiple years of data to enable such analyses. Powerful insights can be derived from the analysis of EMR data, but it is essential to take appropriate steps to understand data provenance and quality prior to analysis to avoid erroneous inferences.

Ethics statement

We received approval from Western University Research Ethics Board for this project (Project ID: 108976; July 31, 2018).

Conflict of interests statement

The authors have no conflicts of interest to declare.

Publication consent

Consent for publication was obtained from DELPHI database administrators. For more information on DELPHI, including data access, see https://www.schulich.uwo.ca/familymedicine/research/csfm/research/current_projects/delphi.html.

Funding statement

This project was funded by a Lawson Health Research Institute Internal Research Fund grant for pilot studies.

References

1. van Velthoven MH, Mastellos N, Majeed A, O'Donoghue J, Car J. Feasibility of extracting data from electronic medical records for research: an international comparative study. BMC Med Inform Decis Mak. 2016 Jul 13;16(1):90. Available from: https://doi.org/10.1186/s12911-016-0332-1.

2. Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. Review: Use of Electronic Medical Records for Health Outcomes Research: A Literature Review. Med Care Res Rev. 2009 Dec 1;66(6):611–38. Available from: https://doi.org/10.1177/1077558709332440.

3. Birtwhistle R, Williamson T. Primary care electronic medical records: a new data source for research in Canada. CMAJ. 2015 Mar 3;187(4):239–40. Available from: https://www.cmaj.ca/content/187/4/239.

4. Muller S. Electronic medical records: the way forward for primary care research? Fam Pract. 2014 Apr;31(2):127–9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3969524/.

5. van Weel C . Longitudinal Research and Data Collection in Primary Care. Ann Fam Med. 2005 May 1;3(suppl 1):S46–51. Available from: https://www.annfammed.org/content/3/suppl_1/S46.

6. Nicholson K, Terry AL, Fortin M, Williamson T, Bauer M, Thind A. Prevalence, characteristics, and patterns of patients with multimorbidity in primary care: A retrospective cohort analysis in Canada. Br J Gen Pract. 2019 Sep;69(686):E647–56. Available from: https://doi.org/10.3399/bjgp19X704657.

7. Williamson T, Green ME, Birtwhistle R, Khan S, Garies S, Wong ST, et al. Validating the 8 CPCSSN Case Definitions for Chronic Disease Surveillance in a Primary Care Database of Electronic Health Records. Ann Fam Med. 2014 Jul;12(4):367–72. Available from: https://www.annfammed.org/content/12/4/367.long.

8. Terry AL, Chevendra V, Thind A, Stewart M, Marshall JN, Cejic S. Using your electronic medical record for research: a primer for avoiding pitfalls. Fam Pract. 2010 Feb;27(1):121–6. Available from: https://academic.oup.com/fampra/article/27/1/121/478208?login=false.

9. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc JAMIA. 2013 Jan;20(1):144–51. Available from: https://academic.oup.com/jamia/article/20/1/144/2909176.

10. Carr H, de Lusignan S, Liyanage H, Liaw ST, Terry A, Rafi I. Defining dimensions of research readiness: a conceptual model for primary care research networks. BMC Fam Pract. 2014 Nov;15:169. Available from: https://bmcprimcare.biomedcentral.com/articles/10.1186/s12875-014-0169-6.

11. Terry AL, Stewart M, Cejic S, Marshall JN, de Lusignan S, Chesworth BM, et al. A basic model for assessing primary health care electronic medical record data quality. BMC Med Inform Decis Mak. 2019 Dec;19(1):30. Available from: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0740-0.

12. Reimer AP, Milinovich A, Madigan EA. Data quality assessment framework to assess electronic medical record data for use in research. Int J Med Inf. 2016 Jun;90:40–7. Available from: https://www.sciencedirect.com/science/article/abs/pii/S1386505616300478#:...:text=A%20six%2Dstep%20data%20quality,%2C%20and%20(6)%20prediction.

13. Stewart M, Thind A, Terry AL, Chevendra V, Marshall JN. Implementing and maintaining a researchable database from electronic medical records: a perspective from an academic family medicine department. Healthc Policy Polit Sante. 2009 Nov;5(2):26–39. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2805138/.

14. Levinson W, Kallewaard M, Bhatia RS, Wolfson D, Shortt S, Kerr EA. 'Choosing Wisely': a growing international campaign. BMJ Qual Saf. 2015 Feb;24(2):167–74. Available from: https://qualitysafety.bmj.com/content/24/2/167.

15. College of Family Physicians of Canada, Choosing Wisely Canada. Family Medicine Thirteen Things Physicians and Patients Should Question. 2020. Available from: https://www.cfpc.ca/CFPC/media/Resources/Pain-Management/Family-Medicine-Thirteen-Things-Physicians-and-Patients-Should-Question.pdf.

16. Shadd J, Ryan BL, Maddocks H, Thind A. Patterns of referral in a Canadian primary care electronic health record database: retrospective cross-sectional analysis. J Innov Health Inform. 2011;19(4):217–23. Available from: https://pubmed.ncbi.nlm.nih.gov/22828576/.

17. Ryan BL, Maddocks HL, McKay S, Petrella R, Terry AL, Stewart M. Identifying musculoskeletal conditions in electronic medical records: A prevalence and validation study using the Deliver Primary Healthcare Information (DELPHI) database. BMC Musculoskelet Disord. 2019 May;20(1):1–8. Available from: https://bmcmusculoskeletdisord.biomedcentral.com/articles/10.1186/s12891-019-2568-2.

18. Maddocks H, Marshall JN, Stewart M, Terry AL, Cejic S, Hammond JA, et al. Quality of congestive heart failure care. Can Fam Physician. 2010;56(12). Available from: https://www.cfp.ca/content/56/12/e432.

19. Christen P, Schnell R. Thirty-three myths and misconceptions about population data: from data capture and processing to linkage. International Journal of Population Data Science. 2023; 8(1). Available from: https://ijpds.org/article/view/2115.

20. de Lusignan S, Liaw ST, Krause P, Curcin V, Vicente MT, Michalakidis G, 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. Contribution of the IMIA Primary Health Care Informatics Working Group. Yearb Med Inform. 2011;6:112–20.

Appendix: Case definitions for indicators

Imaging for lower-back pain

Billing, Diagnosis, Problem, or Health Condition		Procedure
ICD-9 724. (Other and unspecified disorders of back)* *ICD-10 M54.5* (Low back pain)*	AND (within 1 year)	"*xray*" or "*xr*" (free text field)

Antibiotics for upper respiratory tract infections (based on definition constructed by Wu et al. [1]).

Billing, Diagnosis, Problem, or Health Condition		Medication
*ICD-9 460.*or 464.*; ICD-10 J00.* (Acute nasopharyngitis [common cold])*	AND (within 1 year)	Amoxicillin-clavulanate
ICD-9 487.; ICD-10 J11.* (Influenza)*		Cefadroxil
ICD-9 487.; ICD-10 J11.* (Influenza)*		Cephalexin
ICD-9 473.; ICD-10 J32.* (Chronic sinusitis)*		Amoxicillin
		Ampicillin
		Cloxacillin
		Penicillin V
		Pivmecillinam
		Cefaclor
		Cefdinir
		Cefixime
		Cefpodoxime
		Cefprozil
		Ceftibuten
		Cefuroxime
		Gemifloxacin
		Ciprofloxacin
		Norfloxacin
		Ofloxacin
		Levofloxacin
		Moxifloxacin
		Azithromycin
		Clarithromycin
		Erythromycin
		Spiramycin
		Telithromycin
		Sulfamethoxazole
		Sulfamethoxazole-Trimethoprim
		Sulfisoxazole
		Trimethoprim
		Tetracyclines
		Lincosomides
		Nitrofurantoin
		Metronidazole
		Vancomycin
		Clindamycin
		Nitrofurantoin
		Metronidazole
		Vancomycin
		Rifabutin
		Rifampin
		Fidaxomicin
		Fosfomycin
		Linezolid
		Methenamine
		Tedizolid

Chest X-rays and ECGs

Procedure		**Procedure**
"*chest*" (free text field)	AND (within 1 year)	"*xray*" or "*xr*" (free text field) "*ecg*" or "*electrocardiogram*" (free text field)

Screening blood tests

Labs
At least 4 of the following screening blood tests on the same date:

Fasting Glucose
Inclusion terms
'glucose' AND 'fasting'
Exclusion terms
'gest' (gestational)

HbA1c
Inclusion terms
'a1c'

Cholesterol
Inclusion terms
'cholesterol' OR 'hdl' OR 'ldl'

Creatinine
Inclusion terms
'creatinine'
Exclusion terms
'u' (urine) OR 'poc' (point of care)

Complete blood count (CBC)
Inclusion terms
'cbc'

Vitamin B12
Inclusion terms
'b12'

Thyroid stimulating hormone
Inclusion terms
'tsh'

References

1. Wu JH-C, Langford B, Ha R, Garber G, Daneman N, Johnstone J, et al. Defining appropriate antibiotic prescribing in primary care: A modified Delphi panel approach. https://doi.org/103138/jammi2019-0023 [Internet]. 2020 Feb 17 [cited 2021 Sep 5];5(2):62–9. Available from: https://jammi.utpjournals.press/doi/abs/10.3138/jammi.2019-0023.