Using your electronic medical record for research: a primer for avoiding pitfalls

Amanda L Terry a,*, Vijaya Chevendra b, Amardeep Thind c, d, Moira Stewart e, J Neil Marshall f and Sonny Cejic d, g

a Centre for Studies in Family Medicine, Department of Family Medicine, b DELPHI Project, Centre for Studies in Family Medicine, c Department of Epidemiology & Biostatistics, d Department of Family Medicine, e Centre for Studies in Family Medicine, Department of Family Medicine and, f Department of Family Medicine, Schulich School of Medicine & Dentistry, The University of Western Ontario, Ontario, Canada and g Byron Family Medical Centre, London, Ontario, Canada.

*Correspondence to Amanda L Terry, Centre for Studies in Family Medicine, Department of Family Medicine, Schulich School of Medicine & Dentistry, The University of Western Ontario, 100 Collip Circle, Suite 245, London, Ontario, Canada, N6G 4X8; E-mail: aterry4@uwo.ca

Received 5 December 2008; Accepted 10 September 2009

In Canada, use of electronic medical records (EMRs) among primary health care (PHC) providers is relatively low. However, it appears that EMRs will eventually become more ubiquitous in PHC. This represents an important development in the use of health care information technology as well as a potential new source of PHC data for research. However, care in the use of EMR data is required. Four years ago, researchers at the Centre for Studies in Family Medicine, The University of Western Ontario created an EMR-based research project, called Deliver Primary Health Care Information. Implementing this project led us to two conclusions about using PHC EMR data for research: first, additional time is required for providers to undertake EMR training and to standardize the way data are entered into the EMR and second, EMRs are designed for clinical care, not research. Based on these experiences, we offer our thoughts about how EMRs may, nonetheless, be used for research. Family physician researchers who intend to use EMR data to answer timely questions relevant to practice should evaluate the possible impact of the four questions raised by this paper: (i) why are EMR data different?; (ii) how do you extract data from an EMR?; (iii) where are the data stored? and (iv) what is the data quality? In addition, consideration needs to be given to the complexity of the research question since this can have an impact on how easily issues of using EMR data for research can be overcome.

Keywords. Electronic medical records, primary health care, research.

Introduction

In Canada, use of electronic medical records (EMRs) among primary health care (PHC) providers is relatively low.1,2 This is not the case in other countries, such as UK and the Netherlands, where the adoption of EMRs into clinical practice is much higher.3 Increasingly, physicians are receiving remuneration for achievement of targets in preventive care and other quality of care activities, particularly in the USA and the UK.4,5 This has stimulated greater interest in the use of EMRs because they can aid in reporting on these targets.6 It appears that EMRs will eventually become more ubiquitous in PHC. This represents an important development in the use of health care information technology as well as a potential new source of PHC data for research. However, care in the use of EMR data is required.7

Four years ago, researchers at the Centre for Studies in Family Medicine, The University of Western Ontario created an EMR-based research project called Deliver Primary Health Care Information (DELPHI). This project successfully implemented information technology (using one commercial EMR software product) in ten PHC group practices across Southwestern Ontario, Canada. The DELPHI team also created a searchable database, built from de-identified EMR data extracted from these practices.

Implementing this project led us to two conclusions about using PHC EMR data for research: first, additional time is required for providers to undertake EMR training and to standardize the way data are
entered into the EMR; second, EMRs are designed for clinical care, not research. Based on these experiences, we offer our thoughts about how EMRs may, nonetheless, be used for research. Many of the issues raised would be encountered by individual PHC providers and researchers and are not particular to the experience of creating a large research database. This paper is focused towards family physicians who are interested in using their own EMRs for research purposes. This type of research is intended to be shared with colleagues or the broader research community and requires ethics review. There are other elements involved in EMR-based research, such as data linkage, as well as privacy and confidentiality issues, which are not addressed here. We discuss four questions to consider when using EMRs to answer research questions and comment on the issues that may be encountered depending on the complexity of the research question. The focus of this paper is on EMR systems which are currently in place and the implications of these systems for those who are contemplating using EMR data for research; our intent was not to list considerations for selecting an EMR for physicians who are not yet using this technology.

First, we will present the results of a literature review, in which we found seven selected manuscripts which are used in the following sections of this paper. We searched the bibliographic databases PubMed, Medline and Cumulative Index to Nursing and Allied Health Literature for the period 2004–09 using a combination of keywords for family medicine and EMRs. Inclusion criteria included English language articles about human participants. This search yielded 781 citations. A total of 426 duplicate citations were excluded. Citations with no abstracts (editorials, letters, news, reports, commentaries and unavailable = 22), and those unrelated to the topic (48) were excluded. A total of 285 abstracts were placed into categories which included: research using EMR data (134); EMR implementation and adoption studies (59); development/use of EMR data (31); EMRs and patient care (26); data quality (19); EMR development (14); as well as EMR ethics/privacy (2). Slightly less than half used EMRs as a data source (47%), followed by those that focused on the implementation and adoption of EMRs (21%). Next were abstracts that discussed the development and use of EMR data, such as the development of databases or measurement of disease outcomes (11%) and those that explored the impact of EMRs on patient care, for example patient and physician communication or the impact of alerts within the EMR (9%). A smaller proportion of studies explored the quality of EMR data (7%). Studies discussing EMR development, such as tools that work within the electronic record to report on chronic disease management (5%) and those relating to ethics/privacy topics comprised the smallest proportion (1%). The seven studies selected from this group are referenced at the end of each of the following sections.

Using EMR Data for Research

Why are EMR data different?

Because EMRs are designed to support health care provision, they are not structured in a way that facilitates the research process. Five characteristics of the data in EMRs (described more fully below) affect their use for research: (i) providers decide where to put information (uniqueness of use); (ii) information may be entered in free-text form instead of being entered in defined fields or picked from a structured list of medical terms; (iii) providers use different terms for the same information (lack of standardization); (iv) information may not be stored in a way that is readily searchable and (v) data that are not important to clinical care may be missing.

The way that information is entered into the EMR is important. Looking at Figure 1, we see features of an EMR that make it more ‘clinician friendly’; however, these features have implications for using the EMR for research. The family physician researcher will want to evaluate their EMR system already in use according to this schema. First, a computer record generally provides the user with a number of options where the same information can be stored; for example, whether or not a pap test has been done may be found in the in-office examination section, the investigation section or with the laboratory results. This means the family physician researcher must be very familiar with the software and scan all of these possible areas to find the data of interest.

Second, EMR users may prefer to enter free text rather than be limited to picking items from drop-down lists for information such as the problem list, medication and referrals. Free-text data that are extracted from EMRs require coding to aggregate the data into specific variables. Some free-text data may not be searchable at all.

Third, using free text means that variability exists as different entries are made for the same item by an individual over time and across individual patients; for example, ‘type 1 diabetes mellitus’ may appear as ‘DM1’, ‘diabetes’ and dozens of other synonymous text strings, all of which must be either searched for or re-coded into standard categories. Furthermore, abbreviations which are clear to users will require investigation if they are to be used for research; for example, the use of a specific physiotherapist’s name or ‘PT’ should be re-coded as ‘physiotherapist’.

Fourth, medical record software generally provides memo fields, where users enter a long narrative for
the visit note. Extensive use of these large blocks of text rather than coded fields makes automated analysis via queries difficult without a great deal of manipulation. As well, interpretation is a challenge and is time consuming, reminiscent of a classical chart audit. In addition, most computer records allow storage or linking of digitized reports that are stored in the computer. The images are not analyzable; however, the typed text component may be read with optical character recognition software. While free text can be searched, this approach remains practically difficult.

Fifth, clinical software generally makes only a minimal number of fields mandatory because all data are not required for all patients. Therefore, data that may be important for specific analyses are often missing, for example, a date that a referral was made. However, there are ways to adjust for some of these situations. Training may be provided to those who are entering data to ensure terms are entered in a standard way, for example, everyone in the practice agrees to use the term ‘BP’ instead of variations of ‘blood pressure’. Some work can be done to address certain issues, such as recoding synonymous terms into one category. Additionally, elements can be added to software programs, such as the use of a medication database where all medications are named consistently and are updated regularly. Taking these steps can assist in making EMR data more useable for research purposes.

Researchers using large EMR-derived databases have had to address at least some issues of data cleaning and standardization to ensure that complex research questions could be answered properly.\(^{13,14}\) Even in situations where less complex research questions are posed, it is essential for researchers to reflect on the possible impact of these five EMR elements on the data that may be used for their studies.

**How do you extract data from an EMR?**

Once issues in accessing the data as a whole have been addressed, there are five basic options for searching for specific data in an EMR. The availability of these features also depends on the type of EMR in use. Looking at Figure 2, we can see that the first level of data extraction is labelled ‘pre-determined queries’. In this case, users select a query option from the software menu (pre-loaded into the EMR), giving the highest ease of use, but the lowest ability to conduct complex searches. For example, this type of query could produce a report that lists patients by diagnosis seen in the previous month. At the second level, labelled ‘simple customizable queries’, users have some input into the queries to generate reports. For instance, the user would select the patients by diagnosis that were seen within a specific start date and end date. The third level, labelled ‘advanced customizable queries’, allows a greater amount of user input than the second level, often using Boolean logic. The user can, for example, generate a report of patients seen between a specific start date and end date and then select only patients taking medication A or medication B and not taking medication C. The fourth level involves using a special interface to enter structured query language (SQL) commands. The interface, if available at all, does not generally provide full SQL functionality. A user would be able to generate tabulations such as the number of patients by provider, diagnosis, age and sex and who have not been in the office during the 6 months prior to their last visit. The fifth level of data extraction, using database tools, provides the highest level of ability to conduct complex searches and is the most...
challenging to use. Virtually any question can be answered if the data are in the EMR database. The user must have the EMR’s entity relationship diagram that shows how various data files are related to one another. The data dictionary that documents each element in the database with a description and data type is also necessary. As an example, with the database tools, users can generate a list of unique diagnoses, assign a category to each according to a particular research question and apply the categorization back to the EMR data. Therefore, a table of the number of patients and the number of visits by physician per each user diagnosis category can be easily generated.

Going beyond level three would likely require working with the specific EMR vendor to obtain the information that would allow this advanced level of data extraction. Alternatively, the vendor may perform the data analyses. Both options may involve associated additional costs. Extracting and analyzing data using database software (level five) may involve collaborating with other PHC providers, researchers and information technology professionals since it often involves the creation of a separate database for analyses.

Extracting data to answer less complex research questions is possible using the four levels described above. However, with an increasingly complex research question, the fifth level of data extraction will likely be required. Many complex studies from the United Kingdom15–17 and the United States13,18 are based on EMR data which are extracted and then pooled using sophisticated methods to create large researchable databases.

Where are the data stored?
EMR data are generally stored either on-site within the computer server that supports the EMR or off-site by an information technology company. In the latter, patient records and associated data are accessed by PHC providers remotely over a secure network, such as a secure Internet portal. Off-site EMRs require working with the information technology company to fully access the database.

What is the data quality?
The quality of the data in the EMR is an important consideration—poor quality data can negatively affect the results of research studies as well as the functioning of the EMR (e.g. decision support functions may not work properly).19 We begin this discussion of data quality with two central issues. The first issue is for EMRs to better reflect co-morbidities than data based on one code per service (as in a fee-for-service model of remuneration). The complete coding of all presenting problems is a pre-requisite for case ascertainment and for valid comparison of patients among practices where one must compare like with like and avoid comparing more complex patients (with many co-morbidities) with less complex patients (with fewer co-morbidities).

The second key issue is developing criteria for identifying patients who have the condition to be studied. If one is interested in studying patients who have a particular condition for example, we need to have confidence that we can find these individuals in the EMR. There are many places within the EMR that may need to be searched to identify these patients, including problem lists, billing codes, medication lists and physical examination results. There are additional considerations if family physician researchers wish to conduct a multi-practice study. After extracting data in a manner that acquires the same information on patients of each PHC provider, one must manipulate the data to answer the research question. How many diabetic patients have regular HbA1C testing? This question requires more than merely counting the number of HbA1C tests. It requires that one knows who is diabetic, using the same rules, in each practice. This is not as easy as it seems. When the DELPHI project compared patients who had the following four definitions of being diabetic, we found the group of patients did not overlap as much as we had expected. The four definitions were: (i) the patient’s active medication list included a plasma glucose lowering agent; (ii) the patient’s problem list contained diabetes mellitus; (iii) the laboratory list contained more than one result for HbA1C test and (iv) the patient had a diagnosis using the International Classification of Primary Care20 of either T89 (type 1 diabetes) or T90 (type 2 diabetes). Without a robust denominator, one cannot create a proportion of patients with good quality care. Therefore, careful detection of disease cases or conditions is...
crucial. In addition, identifying how long a patient has had a problem and determining the status of the problem are more complex issues than finding an individual with a specific prevalent condition. Researchers need to work to develop processes for solving issues such as the identification of the first diagnosis of a problem in EMR data.

We present four practical examples of reasons for low data quality. One of these is the reliance on potentially inaccurate billing codes as the means to identify groups of patients with the same condition. These codes are used in health administrative data and have been found to need supplementary information before providing accurate groupings. Since EMRs have much more data available, using billing codes only is not using the EMR to its fullest potential, and identifying the supplementary information needed for accuracy in EMR research is an important task for PHC researchers. A second example is incomplete data on topics not completely essential for clinical care, which may therefore be included infrequently by practitioners (e.g. drug identification numbers). A third example is data that are needed to answer a research question for example, socio-economic status, that are not included in the EMR software. This problem requires thorough study of the software’s capability in advance of investing too much research time and energy. Fourth, the use of parallel paper and electronic systems, particularly at the initial stages of EMR implementation, can lead to missing data in the electronic records. Quality control and finding ways to increase the completeness of data overcome the issues just raised.

There are many ways to assess the quality of EMR data. Recent work in the UK suggests that assessments include, at minimum, analyses of comparability (through comparisons with other data sources), completeness and correctness [Hogan and Wagner refer to completeness as ‘... the proportion of observations made about the world that were recorded in the (computer-based patient records) CPR...’ given by a sensitivity value; and correctness as ‘... the proportion of CPR observations that are a correct representation of the true state of world...’ given by a calculation of positive predictive values]. If the quality of the data is unacceptable for research purposes, it may be necessary to improve the completeness and standardization of data entry. Consideration may need to be given to using a classification system with agreed-upon definitions, such as the International Classification of Primary Care.

Both relatively simple and more complex research questions require data of high quality. This is so that the first step of identifying individuals with a specific condition may be made correctly, and second, so that further analytic steps based on these identified individuals (using other EMR data) may be taken with confidence. Care is required in using EMR data for research as variability in levels of quality have been identified.

Conclusion

Family physician researchers who intend to use EMR data to answer timely questions relevant to practice should evaluate the possible impact of the four questions raised by this paper: (i) why are EMR data different? (ii) how do you extract data from an EMR? (iii) where are the data stored?; and (iv) what is the data quality? In addition, consideration needs to be given to the complexity of the research question since this can have an impact on how easily issues of using EMR data for research can be overcome. Despite the challenges associated with using EMR data for research, there are great benefits, such as access to data that are not otherwise available, speed with which analyses can take place and size of the datasets which can be created. This technology is likely to continue to evolve, potentially supporting further research uses of EMR data.

Declaration

Funding: The DELPHI project is not funded by agencies other than those listed below. The DELPHI Project was funded by the Canada Foundation for Innovation; Ontario Ministry of Health and Long-Term Care. Researchers are funded by: Dr Brian W. Gilbert Canada Research Chair (to MS); Canada Research Chair in Health Services Research (to AT); Canadian Institutes of Health Research Strategic Training Program on Transdisciplinary Understanding and Training on Research—PHC (to ALT); Canadian Health Services Research Foundation and the Canadian Institutes of Health Research Post-Doctoral Fellowship (to ALT).

Ethical Approval: The DELPHI project received approval from The University of Western Ontario Review Board for Health Sciences Research Involving Human Subjects.

Conflict of interest: None.

References


