Informatics in Primary Care 2007;15:121–7

Conference paper

Electronic health records: high-quality electronic data for higher-quality clinical research

Mark G Weiner MD
Assistant Professor of Medicine, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

Jason A Lyman MD MS
Assistant Professor of Research, Department of Public Health Sciences, University of Virginia School of Medicine, Charlottesville, VA, USA

Shawn Murphy MD PhD
Assistant Professor of Neurology, Harvard University and Director, Research Patient Data Registry, Massachusetts General Hospital, Boston, MA, USA

Michael Weiner MD MPH
Associate Professor of Medicine, Director, Gero-Informatics, Regenstrief Institute, Inc., Indiana University Center for Aging Research, Indianopolis, IN, USA

ABSTRACT

In the decades prior to the introduction of electronic health records (EHRs), the best source of electronic information to support clinical research was claims data. The use of claims data in research has been criticised for capturing only demographics, diagnoses and procedures recorded for billing purposes that may not fully reflect the patient's condition. Many important details of the patient's clinical status are not recorded.

EHRs can overcome many limitations of claims data in research, by capturing a more complete picture of the observations and actions of a clinician recorded when patients are seen. EHRs can provide important details about vital signs, diagnostic test results, social and family history, prescriptions and physical examination findings. As a result, EHRs present a new opportunity to use data collected through the routine operation of a clinical practice to generate and test hypotheses about the relationships among patients, diseases, practice styles, therapeutic modalities and clinical outcomes.

This article describes the clinical research information infrastructure at four institutions: the University of Pennsylvania, Regenstrief Institute/Indiana University, Partners Healthcare System and the University of Virginia. We present models for applying EHR data successfully within the clinical research enterprise.

Keywords: biomedical research, hospital information systems, medical record systems, computerised

University of Pennsylvania: the PICARD System

For many years, investigators at the University of Pennsylvania School of Medicine were able to conduct research using electronically stored data collected through the day-to-day operation of the University of Pennsylvania Health System. However, as is typical of many health systems, much of the early data related to administrative and billing activities. Most outpatient diagnosis data were recorded by physicians at the end of a clinical encounter, while coding professionals recorded diagnoses based on the clinical

documentation of an inpatient admission. Demographics were recorded at the time of patient registration. Despite many recognised shortcomings, these data enabled research that explored patterns of ambulatory and inpatient diagnoses that were predictive of length of stay, mortality and readmission. As with billing systems, electronic laboratory systems have a long history at the University of Pennsylvania, though their use has been optimised for reporting on single patients, not cohorts required for research. Research that spanned administrative and clinical laboratory data, therefore, was cumbersome because of both the difficulty in extracting comprehensive laboratory information on populations and the need to deal with two separate data centres regarding the information need. Even research questions that spanned the inpatient and outpatient environments were challenging because the information systems for these environments, and the staff that maintained them, were entirely distinct. The data collection process was inefficient, requiring the investigator to carry out substantial data merging and refinement, and resulting in the transfer of far more data than were needed in the final analysis.

In 1997, to address these inefficiencies in data acquisition, we began work on the Pennsylvania Integrated Clinical and Administrative Research Database (PICARD), and developed an administrative infrastructure to provide a single point of contact for investigators to address their research information needs. PICARD started as a refinement of existing outpatient billing data systems, in which the basic record unit was an invoice, into a relational data model where each record represented a different outpatient visit. Information on inpatient encounters was integrated in the data model to enable both summary overview of inpatient admissions as well as daily detail of clinical activity, including diagnoses and procedures. Laboratory investigations were included within the same data model. As PICARD grew in breadth and longitudinal scope, the research questions it could address became more sophisticated. Cohorts could be defined initially by the presence of diagnoses and then refined by a pattern of laboratory values and other clinical activity over time.

Over the ensuing years, the Health System has implemented a number of electronic systems to support its clinical mission and these data have also been incorporated into PICARD. The important additions included data from our outpatient EHR and our inpatient order-entry/results reporting system. With these systems, investigators have access to important clinical details collected at the point of care, including vital signs, medications ordered and social history. These details help to create more homogeneous cohorts of patients based on true clinical status, or provide important covariate to adjust for imbalances in clinical characteristics. The comprehensive, longitudinal data also enables tracking of health utilisation and changes in clinical status over time.

PICARD currently tracks information on over 1.8 million patients seen in our three inpatient hospitals and our primary care and subspecialty outpatient practices since 1997, representing over 25 million encounters. Over 46 million diagnoses have been assigned and more than 153 million laboratory tests have been recorded. The scope of the data on outpatient visits and inpatient admissions includes all patient demographics, location of the encounters and participating physicians, as well as diagnoses assigned during each encounter and charges and reimbursements for all procedures performed. Among the 184 000 patients seen within practices using the electronic medical record, we have additional discrete details about medication prescribing, social history (including smoking, alcohol and drug use) and vital signs including weight, height, blood pressure, pulse and respiratory rate.

The clinical setting under which the data are collected requires several caveats when interpreting the meaning of data for research purposes. Under a research protocol, subjects are followed at regular intervals and receive testing that is the same for all participants. In the clinical setting, patients present for outpatient visits at scheduled intervals and receive specialised testing in a manner determined by the complexity of the individual patient's array of comorbidities, or when the patient is not feeling well. As a result, there may be systematic bias in the volume of data favouring patients with more underlying disease. Functional status surveys taken at the time of a visit to a physician might reflect the patient's acute medical problem, rather than the patient's health on most other days of the year. Data on a single patient might contain conflicting or ambiguous concepts, as patients see multiple physicians over time. Patients might receive portions of their health care at other institutions, leaving gaps in the researchers' understanding of health utilisation and ancillary test results.

Despite these limitations, the comprehensive nature of clinical practice databases like PICARD offers several advantages over older, mostly administrative data sets used for research. The value of these databases lies in targeted recruitment for clinical trials and for primary data collection for health services and epidemiological research. They can be used for hypothesis testing and generation. While randomised controlled trials remain the gold standard for the assessment of efficacy of interventions, databases such as PICARD can be used to extend the generalisability of clinical trial results to other populations, or to confirm if older results are still valid, and could help provide insight into a research question if a formal trial would be prohibitively expensive or unethical to conduct.

Regenstrief Institute: Regenstrief Medical Record System (RMRS)

At the Regenstrief Institute, on the campus of the Indiana University School of Medicine in Indianapolis, the RMRS has been developed to provide electronic medical records for Wishard Health Services.[1] With computerised order entry, the RMRS contains more than 20 million physicians' orders for three million patients and includes data from pharmacy, diagnostics, procedures, narratives and radiology.

In the pursuit of interoperable systems to improve clinical care and research, a wide area network was created to share data among multiple local teaching and community hospitals. This grew from the Indianapolis Network for Patient Care and Research into the Indiana Network for Patient Care, a local health information infrastructure targeting five major hospital systems, as well as public health departments, Indiana Medicaid and RxHub.[2–4] With patients' permission, this system allows physicians to view data from multiple hospitals within the network. The network also delivers diagnostic test results and other documents to most medical practices in the area.

The electronic system can be used in a variety of ways to conduct research.

- It has been used to conduct many secondary data analyses in a retrospective fashion.
- It can be used as a source of data to link to other sources, such as Medicare or Medicaid data.
- It can be used to track key clinical measures in prospective studies.
- It has been used for internal research, such as local monitoring of providers' practices for quality improvement.
- The system itself can be modified and used to deliver interventions, such as clinical decision support.

Several important components or adjuncts have facilitated these types of research. First, originating in primary care, a practice-based research network has been created.[5] Funded by the Agency for Healthcare Research and Quality, this network has a business agreement with the academic medical practice to recruit subjects for research. It provides research associates in the network with access to clinical data for assessing eligibility for studies. With this organised network, more than 8000 patients have been recruited into studies. Second, a data management team has been formed to handle many data processing tasks for research and local quality improvement.[6] This team responds to authorised requests for data, develops query syntax based on requests or study designs, executes queries, examines data extractions for completeness and validity, and delivers data to analysts for further processing.

Finally, data query tools have been developed to allow investigators to retrieve data directly from the system. In the latest rendition, evolved from a pathology project to give researchers limited access to de-identified clinical data, users can generate queries of a broader set of clinical data via a graphical interface.[7] The user undertakes a three-step process to generate a query and retrieve results. First is the definition of a cohort of interest. The user selects variables of interest from the large array contained in the clinical repository. Specific variables, such as respiratory rate, and specific values, such as 'greater than 20', can be selected, as well as any particular demographics. Next, the data elements needed for the report are chosen. These might be related to encounters, blood test results or many other clinical parameters. Next, the analysis plan is defined. Examples are cross-tabulation, regression analysis and survival analysis. The system then executes the query and provides the output.

We find that a few general rules of thumb apply to most studies undertaken through use of the system. It helps to form a study team and plan an organised approach to retrieving, managing and analysing data, rather than using a more 'spur of the moment' method. This work takes dedicated time and the presence of an electronic medical records system should not be taken to mean that the analysis will become easy or straightforward. In contrast, as with other forms of data used for research, these require validation and special study for completeness and accuracy. For example, although our system collects nearly all clinical data in our environment, a recent study showed that data from one source missed about 40% of documentation of patients' vaccination and mammography, simply because of the ways in which patients obtain health care from multiple sites across institutions.[8]

Such a rich data network and supportive infrastructure provide nearly limitless opportunities to conduct research, using tools from clinical sciences, health services and increasingly even basic science. Conducting clinical or health services research in primary care stems naturally from this resource and provides capacity to undertake a large number of projects.[9] Our institution has used this informatics programme to conduct work especially related to informatics interventions, quality, efficiency and resource utilisation associated with health care.

The Clinical Data Repository (CDR) at the University of Virginia Health System, Charlottesville, VA

Introduction

At the University of Virginia Health System (UVaHS) we have a decade of experience with data warehousing in an academic health centre, with a focus on supporting clinical and health services investigation.[10] Using a custom-developed, web-based user interface, our researchers can create *ad hoc* flexible queries of retrospective patient data and view a wide variety of anonymised reports. Individuals with appropriate authorisation can work directly with the CDR team to request identifiable patient data as needed. Our database contains data on over 850 000 patients and five million encounters, spanning all care settings affiliated with the University of Virginia Health System. There are multiple issues that affect the extent to which researchers are successful in using the CDR to conduct their projects. Data availability, data format, data accuracy and user interface issues are several factors that must be considered for anyone developing, modifying or using data warehouses for clinical investigation. For purposes of example we describe a recent scenario in which a junior researcher came to us for help in using the CDR to explore the association between the use of specific antibiotics and the development of *C. difficile* enterocolitis in patients hospitalised at UVaHS.

Data sources/availability

When considering the use of a CDR for a specific research project, the first question our users often ask relates to the specific data contents of our system. This is typically more complicated than they initially recognise. Ideally, they need not only the outcome data of interest, but also the necessary means for identifying patients (often based on diagnosis, procedures, demographic factors and so on) and, in some cases, data to identify any important confounding factors. The CDR contains administrative data (coded diagnoses and procedures, utilisation data and demographic information, all captured for billing purposes), data on medications administered within UVaHS, clinical laboratory and microbiology results and mortality data from the Virginia Department of Health. We lack narrative data such as discharge summaries, pathology reports and progress notes. Medication prescribing data is also currently unavailable, pending implementation of our outpatient EHR. For our researcher in the current example we needed access to inpatient medication data, microbiology results, demographic information and a way to limit our query purely to inpatient cases.

Data format

Data format issues also must be considered. Much of the information in the medical record is unstructured, narrative data. Textual data abounds, even in potentially unexpected places like clinical laboratory results (such as urine protein) and microbiology results. Such data are difficult to use reliably in queries for several reasons, including misspellings, synonyms, homonyms and negation, to name a few. For our example project, medication data were coded using an internal system that, while cumbersome, allowed us to identify patients who had received the antibiotics of interest. The microbiology results, used for identifying patients with a positive *C. difficile* test, were in text format. For this particular test, however, positive results were always represented using the same text, enabling us to successfully identify these cases in our system.

Data accuracy

Data accuracy varies tremendously, based on multiple factors. Diagnoses encoded in administrative data, for example, are often less sensitive and specific than we would like, though this varies depending on the clinical concept being represented. Clinical laboratory results, captured directly from laboratory information systems, tend to be quite accurate. Medication administration data, collected for both billing and clinical purposes, also tend to be of higher quality than administrative data. Both of these latter categories were used in our example project as the prime data sources for identifying patients, so we felt reasonably confident in the use of the CDR data to support the researcher's effort.

User interface

One of the more unusual aspects of our CDR is its web-based user interface, which allows local authorised users direct access to a powerful query-generating tool and includes a variety of aggregate and detailed reports. While our ultimate goal is for users to be able reliably and independently to use the interface to complete their projects, several factors make this goal elusive. Users are frequently not familiar with the underlying coding systems used at UVaHS. The flexibility provided to allow robust queries also makes for a complex interface. For the project under consideration, the researcher was fortunately able to do

the vast majority of the data retrieval herself using our web interface, though our team worked with her to determine the best way to identify the cases of interest.

Conclusion

The CDR, developed specifically to support the research mission of our academic medical centre, has been used to support hundreds of projects in recent years. As the implementation of our UVaHS-wide EHR continues, we look forward to receiving additional, more clinically orientated data that will enrich our system and increase its utility to our researchers. While such information will improve the CDR from a 'data availability' standpoint, the challenges around data formatting and accuracy, and providing a user-friendly user interface, will remain.

Partners Health Care System: Research Patient Data Registry (RPDR)

The Research Patient Data Registry (RPDR) at Partners Healthcare serves as a data warehouse that integrates clinical, administrative and research data from many data sources for the primary purpose of supporting research. Researchers can access the RPDR database using a web-based query tool, designed as an integral part of the project and accessible from any computer workstation on the private Partners Healthcare intranet.[11] Authorised users may query against RPDR data for aggregate totals and, with proper International Review Board (IRB) approval, may obtain specific patient-identifiable clinical data. This capability allows researchers to quickly obtain information that can be critical for winning corporate and government sponsored research grants, and easily gather data on patients identified for research studies. Security and confidentiality are an integral part of the project and the RPDR brings clinical information to researchers' fingertips while controlling and auditing the distribution of patient data within the guidelines of the IRB.[12]

The RPDR database is composed of over four million patients and 900 million coded records from patient encounters, laboratory investigations and results and other medical care. Each coded event is represented as fact in the database and in turn associated with other important contextual information. The scope of the RPDR includes not only patient demographic data, diagnoses, procedures, pharmacy data, inpatient and outpatient encounter information, provider information and laboratory data, but also data from the longitudinal EHR (LMR). The RPDR has over 1350 users throughout the Partners' Healthcare system. Since its inception in 2002, a total of 2155 identified data sets containing a total of over 10 million patient records have been returned to RPDR users. Estimated money in 2005, funded by sponsors to grants that were critically dependent on the RPDR, ranged from US$20.7 million to US$30.7 million, and their total funding ranged from US$94 million to US$136 million.[13]

To further increase this return on investment, the RPDR team has been focusing on four 'building block' applications that will further develop the capabilities of medical records research.

1. Bayesian inference engine

As the number of sources for clinical data on a patient increases, a new problem arises. The new problem is that conflicting data are often found within the database on the patient. Bayesian inference can be used automatically to reduce conflicting and scattered observations into fundamental atomic concepts regarding a patient. For example, a code might be assigned to a patient from several sources indicating that a patient has a disease such as diabetes. However, some sources may indicate the patient has type I diabetes, while others indicate the patient has type II diabetes. Since these two types of diabetes are virtually exclusive, it is clear that one of the sources is in error. A determination of the true diagnosis can be estimated by assigning a prior probability to each source as to how often it contains correct information. For example, data from an endocrine clinic would be assigned a high value. One then uses these probabilities to calculate the likelihood of each diagnosis.

2. Predictive modelling

Predictive modelling in health care can be applied to a variety of problems such as finding high-risk patients, thus allowing early intervention. This serves towards both cost containment and decreasing medical errors. For example, in treating asthma there are several variables that predict those patients that will have another severe asthma attack within a given period of time. Variables such as smoking status, age, the number of prior attacks and results of pulmonary function tests might be shown to predict the likelihood of a new attack. Preventive treatments can then be focused upon these patients to thwart such attacks.

3. Clinical trials performed *in-silico*

Performing an observational phase IV clinical trial is an expensive and complex process that can be potentially modelled in a retrospective database. This application would allow a formalised way of discovering new knowledge from medical databases in a manner that is well accepted by the medical community. However, fundamental problems complicate this approach:

- patients drift in and out of the system. Sophisticated statistical models using adequate control populations are necessary to compensate
- confounding variables may not be coded in the database. Sophisticated natural language processing might be needed to extract the confounders from textual reports in order to allow confounders to be resolved where they cannot be found in coded data
- most clinical databases do not distinguish between the patient known not to have a disease and the disease not being recorded for that patient.

4. Finding correlating relationships within data

Unsupervised techniques using Relationship Networks and Mutual Information algorithms can generate hypotheses from observed correlations in the data. The database can be watched to automatically pull out new correlations that are found between diseases, medications and laboratory values. A correlation means that two or more events are occurring in a temporally related fashion above a given signal/noise ratio. It does not imply that the events are causally related or that they have any medical significance. Therefore this method can be used to suggest new hypotheses, but these then need to be investigated manually.[14]

ACKNOWLEDGEMENTS

Michael Weiner is supported by the National Institute on Aging, 5K23AG020088. Shawn Murphy is supported in part by the NIH Roadmap for Medical Research Grant U54LM008748.

This paper was presented by the authors as a Scientific Panel of the AMIA Primary Care Informatics Working Group at the American Medical Informatics Association Annual Symposium in Washington DC in November 2006.

REFERENCES

1 McDonald CJ, Overhage JM, Tierney WM *et al.* The Regenstrief Medical Record System: a quarter century experience. *International Journal of Medical Informatics* 1999;54:225–53.
2 Overhage JM, Tierney WM and McDonald CJ. Design and implementation of the Indianapolis Network for Patient Care and Research. *Bulletin of the Medical Libraries Association* 1995;83:48–56.
3 McDonald CJ, Overhage JM, Barnes M *et al.* The Indiana network for patient care: a working local health information infrastructure. An example of a working infrastructure collaboration that links data from five health systems and hundreds of millions of entries. *Health Affairs (Millwood)* 2005;24:1214–20.
4 Biondich PG and Grannis SJ. The Indiana network for patient care: an integrated clinical information system informed by over thirty years of experience. *Journal of Public Health Management and Practice* 2004;(Suppl): S81–6.
5 Agency for Healthcare Research and Quality. *AHRQ Supports 19 Primary Care Practice-based Research Networks.* 2000. www.ahrq.gov/news/press/pr2000/pbrnspr.htm
6 Murray MD, Smith FE, Fox J *et al.* Structure, functions, and activities of a research support informatics section. *Journal of the American Medical Informatics Association* 2003;10:389–98.
7 National Cancer Institute. *Shared Pathology Informatics Network.* www.cancerdiagnosis.nci.nih.gov/spin/
8 Weiner M, Quwatli Z, Perkins AJ, Lewis JN and Callahan CM. Limitation of a single clinical data source for measuring physicians' performance on quality indicators. *Journal of the American Geriatrics Society* 2006; 54:1256–60.
9 Chaudhry B, Wang J, Wu S *et al.* Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. *Annals of Internal Medicine* 2006;144:742–52.
10 Einbinder JS, Scully KW, Pates RD, Schubart JR and Reynolds RE. Case study: a data warehouse for an academic medical center. *Journal of Healthcare Information Management* 2001;15:165–75.
11 Murphy SN, Gainer VS and Chueh H. A visual interface designed for novice users to find research patient cohorts in a large biomedical database. *Journal of the American Medical Informatics Association Symposium* 2003; (Suppl):489–93.
12 Murphy SN and Chueh H. A security architecture for query tools used to access large biomedical databases. *Journal of the American Medical Informatics Association Symposium* 2002;(Suppl):552–56.
13 Nalichowski R, Keogh D, Chueh H and Murphy SN. Calculating the benefits of a research patient data repository. *Journal of the American Medical Informatics Association Symposium* 2006;(Suppl):1044.
14 Butte AJ and Kohane IS. Unsupervised knowledge discovery in medical databases using relevance networks. *Journal of the American Medical Informatics Association Symposium* 1999;(Suppl):711–15.

CONFLICTS OF INTEREST

None.

ADDRESS FOR CORRESPONDENCE

Mark Weiner MD
Assistant Professor of Medicine
University of Pennsylvania School of Medicine
423 Guardian Drive – Rm 1116
Philadelphia, PA 19104
USA
Tel: +1 215 898 5721
Email: mweiner@mail.med.upenn.edu

Accepted April 2007