Spring 5-4-2019

A Beginner's Guide to Research Using Electronic Health Record Data

Jenenne A. Geske
University of Nebraska Medical Center

Tell us how you used this information in this short survey.
Follow this and additional works at: https://digitalcommons.unmc.edu/etd

Part of the Health Information Technology Commons

Recommended Citation
https://digitalcommons.unmc.edu/etd/372

This Thesis is brought to you for free and open access by the Graduate Studies at DigitalCommons@UNMC. It has been accepted for inclusion in Theses & Dissertations by an authorized administrator of DigitalCommons@UNMC. For more information, please contact digitalcommons@unmc.edu.
A BEGINNER'S GUIDE TO RESEARCH
USING ELECTRONIC HEALTH RECORD DATA

by

Jenenne A. Geske, PhD

A THESIS

Presented to the Faculty of the
University of Nebraska Graduate College in
Partial Fulfillment of the Requirements
for the Degree of Master of Science

Biomedical Informatics Graduate Program

Under the Supervision of Professor James McClay

University of Nebraska Medical Center
Omaha, Nebraska

May, 2019

Advisory Committee:
Jane Meza, PhD
Dario Ghersi, MD, PhD
ABSTRACT

Since the passage of the American Recovery and Reinvestment Act in 2009, the use of Electronic Health Records (EHRs) in the healthcare system has increased substantially. Accompanying this surge in EHR usage is a surge in healthcare data and increased opportunities to improve our understanding of health care through research using these data. The use of EHR data for research has many benefits, limitations and considerations. Using data that was originally intended to facilitate billing, insurance, and maintenance of clinical records for research can be fraught with challenges, but they can also be a rich source of information. This paper addresses some of these benefits and challenges, along with additional considerations, including ensuring the best quality data, selecting a good study design, tailoring research questions and queries to available data, and understanding ethical issues in using patient data for research. Researchers should develop a clear understanding of the pitfalls inherent in EHR research before beginning a project. As is the case with most research, many of the drawbacks can be reduced with careful preparation, formulation of a research question, procedures and data management. Appendices include a Flow Chart detailing the EHR research process, and a User's Guide for UNMC's deidentified electronic health records database.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>i</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>ii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>iii</td>
</tr>
<tr>
<td>CHAPTER 1: INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>CHAPTER 2: BENEFITS OF EHR RESEARCH</td>
<td>2</td>
</tr>
<tr>
<td>CHAPTER 3: CHALLENGES TO EHR RESEARCH</td>
<td>4</td>
</tr>
<tr>
<td>CHAPTER 4: CONSIDERATIONS IN DESIGNING EHR RESEARCH</td>
<td>6</td>
</tr>
<tr>
<td>Formulating a Research Question</td>
<td>8</td>
</tr>
<tr>
<td>Identifying a Population or Patients of Interest</td>
<td>9</td>
</tr>
<tr>
<td>Selecting Variables</td>
<td>10</td>
</tr>
<tr>
<td>Data Cleaning</td>
<td>11</td>
</tr>
<tr>
<td>CHAPTER 5: ETHICAL CONSIDERATIONS</td>
<td>16</td>
</tr>
<tr>
<td>CHAPTER 6: CONCLUSION</td>
<td>17</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>19</td>
</tr>
<tr>
<td>APPENDIX A: Questions to Evaluate the Quality of Retrospective Data Studies (from Motheral et al., 2003)</td>
<td>24</td>
</tr>
<tr>
<td>APPENDIX B: Using the Deidentified UNMC i2b2 Database for Research</td>
<td>27</td>
</tr>
<tr>
<td>APPENDIX C: The Medical Records Research Process flow chart</td>
<td>38</td>
</tr>
</tbody>
</table>
LIST OF TABLES

TABLE 1: DATA CLEANING ISSUES BY DATA TYPE.................................................................12
TABLE 2: TECHNIQUES FOR ASSESSING DATA....................................................................14
CHAPTER 1: INTRODUCTION

Since the passage of the American Recovery and Reinvestment Act in 2009, the use of Electronic Health Records (EHRs) in the healthcare system has increased substantially. The percentage of outpatient physician offices that has adopted the use of an EHR more than doubled, from 42% in 2008 to 87% in 2015.\(^1\) Additionally, 80% of non-federal acute care and small hospitals have adopted their use, while almost 100% of all federal acute care hospitals have.\(^2\) Accompanying this surge in EHR usage is a surge in healthcare data and increased opportunities to improve our understanding of health care through research using these data. In fact, improving the ways in which EHRs can be used for research is itself a growing area of research. For example, the National Institutes of Health (NIH) has provided funding to develop a framework that facilitates the process of querying and extracting data from EHRs. Tools using this framework, the Informatics for Integrating Biology and the Bedside (i2b2), initially housed at Partner's Health Center in Boston, are made available for use free of charge.\(^3\)

The use of EHR data for research has many benefits, limitations and considerations. Using data that was originally intended to facilitate billing, insurance, and maintenance of clinical records for research can be fraught with challenges, but they can also be a rich source of information.\(^4,5\) This paper addresses some of these benefits and challenges, along with additional considerations, including ensuring the best quality data, selecting a good study design, tailoring research questions and queries to available data, using appropriate data analysis methods and understanding ethical issues in using patient data for research.
CHAPTER 2: BENEFITS OF EHR RESEARCH

Using EHRs for research purposes offers many benefits when compared to more traditional study designs. They allow researchers to follow patients longitudinally, study safety of various treatments and procedures, and examine the effectiveness of treatments.\textsuperscript{6-10} In addition, they are useful for examining geographic distributions of disease and other characteristics. They are particularly useful for examining stigmatized or rare conditions, in which it might be difficult to identify and recruit patients for a randomized study.\textsuperscript{4,5} Because EHR research accesses existing patient records, often thousands or hundreds of thousands of patients within a system can be considered as research subjects\textsuperscript{3,5,11,12}.

In a related vein, EHR research can provide a great return on investment. It is cost-effective and time-efficient\textsuperscript{3-5,11} Medical records can provide information that is similar to the type of data that might be collected in a non-interventional trial, at a greatly reduced cost.\textsuperscript{12} Most of the cost has been absorbed by the health-care system with the installation and maintenance of the EHR. Additional costs come from salaries and server space.\textsuperscript{3} However, there is less need for a large research staff to recruit subjects and collect data, and few, if any, costs for supplies, etc. In addition, costs associated with repeated office visits and/or assessments of subjects are reduced or eliminated with EHR research.\textsuperscript{9} Although manual coding of information may contribute to delays, EHR research can usually provide timely access to data. There are fewer delays of the type that result from manual data collection and entry. Therefore, information on events that impact patient care, such as safety and treatment options, can be assessed more quickly.\textsuperscript{9}

Although there are some ways in which EHR data can be unreliable (see Challenges to EHR Research, below), there are other ways in which EHR research can result in reduced measurement error. For example, research that relies on self-reported data from patients or other study subjects can often be affected by recall bias, while most EHR data do not have this source
of error. Similarly, EHR data do not typically have biases due to social desirability or Hawthorne effects (changes in a subject's behavior resulting from the fact that they know they are being observed). Although some EHR data are manually entered and are subject to data entry or transcription errors, other data points are automatically entered, reducing the opportunity for errors.

EHRs can be instrumental in the conduct of Comparative Effectiveness Research (CER), which uses data to identify best practices for clinical care. The large samples, potential for long-term follow-up, and diverse samples that can be obtained using EHR data can overcome many traditional barriers to CER, and the relative speed with which data can be gathered from an EHR in comparison to a prospective trial can ensure more timely and accurate decision support for clinicians.

Finally, research using EHRs can be an important supplement to data obtained using more traditional research methods. Randomized controlled clinical trials (RCTs) are still commonly accepted as the gold standard in healthcare research, and for good reason. They are prospective in nature, allowing for a study design that is specifically tailored toward the research question of interest. The randomization of participants to treatment condition theoretically ensures that confounding factors are evenly distributed among all arms of the study and reduces self-selection and researcher bias. Interventions are tightly controlled and their effects can be statistically isolated.

As a supplement to RCTs, EHRs can be used to examine the feasibility of a planned RCT. They are particularly useful in helping to identify a study cohort. For example, they can be used to pre-screen a patient panel to identify patients who meet inclusion criteria. This screening can provide information on possible sample size, and can be used to determine which subjects to approach for participation in a prospective study. EHR data can be used to monitor or facilitate
the delivery of clinical interventions. They can also provide information about rates of events, efficacy of treatments, and patient behavior that can be used to inform experimental designs.

EHR data can be correlated with randomized, controlled trial data to confirm the experimental effect in a non-randomized, real-world population. If a randomized controlled trial, for example, examines two treatments in order to determine which is better for all qualifying patients, an EHR study can then assess how the better treatment works when applied in uncontrolled conditions to representative patients.

Often, data in EHRs come from multiple sources, providing the opportunity to triangulate findings and identify gaps in patient care. In many cases, data in EHRs are recorded at a more granular level, increasing the precision of the data, and the varying sources and granularity of the data may allow the researcher to analyze the data in different ways.

CHAPTER 3: CHALLENGES TO EHR RESEARCH

As previously alluded to, EHR data may have some data quality issues, and EHR methodology has some inherent challenges. Researchers must consider the potential for bias at every step of the research process. One of the greatest challenges to research using EHR data is incomplete data, which calls into question any interpretations or conclusions that arise from it. For example, patients may be lost to follow-up, resulting in missing data points. Researchers rarely know the reason that these patients are lost. Did they conclude their care? Move to another region? Change insurance providers? Receive care from another provider outside of the system? The interpretation of results may vary depending on the reasons patients are lost.

Inconsistent procedures across providers, clinics, departments and/or systems may also contribute to incomplete EHR data. Some physicians’ offices, providers, etc. may record
specific data points, while others may not. Some may record data in structured, searchable fields. Others may enter the same information in an unstructured area of a chart,\textsuperscript{5,19,20} which may not be as easily searched. Natural language processing tools may facilitate analysis of unstructured data,\textsuperscript{3} but gathering data from text fields such as "notes" can be labor intensive and, given the limited amount of data one might recover, may not be worth the time and effort.\textsuperscript{19}

Incomplete data may also occur because the desired data is simply not recorded in an EHR.\textsuperscript{20} For example, patient satisfaction measures or past patient addresses may not be included in the medical record, but may be of interest to a researcher.\textsuperscript{5} Research that examines historical records may also be problematic if data that were gathered before the implementation of the EHR are not be readily available.\textsuperscript{20} Similarly, poorly designed EHRs and changes in EHR vendors or in the data structure of an existing EHR may make it difficult to identify and locate data.\textsuperscript{19}

Identifying the appropriate search terms and fields may be difficult in EHR research. It is not uncommon for data to be recorded using multiple terms for the same concept over time or across providers and/or patients, particularly in unstructured fields.\textsuperscript{5} One research group, for example, examined "notes" fields in 465 charts of patients with otitis media and found 278 different ways in which the provider indicated that the child had a temperature greater than 102.0 F.\textsuperscript{23} Indication of type 1 diabetes mellitus may be recorded as DM1, juvenile diabetes, insulin-dependent diabetes, or by its ICD-9 or ICD-10 disease classification codes.\textsuperscript{20}

Similarly, a medical record may have multiple fields for the same information.\textsuperscript{4,5,9,20} Researchers may need to search physical examination results, problems lists, medication lists, and billing codes to identify patients of interest. For example, in order to identify patients for whom a pap test was done, researchers may need to look in an in-office examination section, lab results, and or an investigation section.\textsuperscript{20} Consequently, researchers must understand all possible locations in which the desired data could be stored, and be diligent in identifying the search term(s) they want to use.
Billing codes that are entered into an EHR may not be accurate, making it challenging to identify patients of interest with a specific disease or condition. Researchers would be wise to identify additional sources of information to verify findings.\textsuperscript{20,21} One researcher determined that in more than half of medical record studies (55%), the EHR data were supplemented with other sources of data.\textsuperscript{4} This process introduces new challenges in that it may not always be easy to match data across multiple sources.

Another challenge in using EHR data is that many data points may have multiple measures for a single patient.\textsuperscript{9} For example, weight is recorded at every office visit, and it changes over time. If a researcher is interested in a patient's weight, they must determine which of the many measurement points they would like to use. Would they prefer the first recorded weight for each patient? The most recent one? An average of all of the measures? Or are they interested in change over time? If that is the case, then all instances of weight are probably important to the researcher. The researcher must consider these questions for each variable obtained from the EHRs.

\textbf{CHAPTER 4: CONSIDERATIONS IN DESIGNING EHR RESEARCH}

Research using EHR data generally falls under the observational research umbrella. Within this category, EHR research can consist of multiple designs and serve many purposes. EHR research is widely used in epidemiologic studies to determine incidence or prevalence.\textsuperscript{9} Studies of rare disease can be conducted using case control designs. Cohort or cross-sectional studies can be conducted for more prevalent diseases or conditions.\textsuperscript{4} Additionally, EHRs are useful for safety surveillance and for comparative effectiveness research.\textsuperscript{9} It should be noted that observational studies, by design, are able to demonstrate associations, but cannot prove causation.\textsuperscript{3}
Because of benefits and challenges integral to EHR research, researchers must familiarize themselves with the data. As previously mentioned, the data in most EHR systems was not originally collected to answer a research question; rather, it was collected to chart patients' health status and to appropriately bill for care. Thus, researchers cannot control which data were collected, how they were collected, or how they were recorded. It becomes very important to acknowledge the biases inherent in this type of research and to determine ways in which to reduce this bias. Researchers must understand how and why the data were collected, and whether and how any further processing was done to the data. They must also ascertain how and where the data are stored within the system. They should have a clear understanding of how the data are measured or observed, the meanings or unit of measurement of any quantitative values, contextual information, when the data were collected and the level of detail recorded.

Additionally, this type of research must be tailored toward the information that is available. In most prospective research designs, a research protocol can be developed that specifies exactly when, where, from whom, what type and how data collection will occur to best answer the research question. However, EHR research generally requires the research question to be tailored to match the existing data.

When designing an EHR research project, there are many factors to consider at each step of the process, from formulating a research question to analyzing data. The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) published a checklist of data, method and conclusion considerations when conducting secondary analyses of existing data (see Appendix 1). This was followed up with a task force consisting of experts in medicine, epidemiology, public health, biostatistics, public health, health economics and pharmacy sciences who published a series of papers targeted toward "recommending good research practices for designing and analyzing retrospective databases."
Formulating a Research Question

Once a research idea begins to emerge, a key first step is to develop a study objective and formulate an effective scientific research question. Researchers must clearly understand the phenomenon of interest, identify the population of interest, define the variables in terms of the information available in the medical record (e.g., how is "improvement in blood pressure" defined) select the relevant data points and determine the best data sources.

While it may not be applicable in all situations, the PICOT strategy may be useful in articulating and clarifying many questions. This strategy offers a framework for structuring a precise research question by specifying the:

(P) Patients, Problem or Population of interest

(I) Intervention to be studied

(C) Comparison or Control condition

(O) Outcome of interest

(T) Timeline

Using this framework aids in clearly articulating the specific components of the research project and can help the researcher begin to think about his or her research question in terms of the variables that must be located within the health records.

For example, a hospitalist and researcher may begin to wonder about the number of fall injuries that are occurring in his patient population. He could ask a general question, "What preventative measure should I use to prevent injuries in my patients?" However, in order to facilitate the data search process, he could formulate his research question using the PICOT framework:
(P) Nebraska Medicine inpatients who have been diagnosed with osteoporosis

(I) Hip protectors alone

(C) On Bisphosphonates alone prior to admission

(O) Injuries from falls

(T) During their stay in the hospital

Thus, the research question becomes, "During their stay in the hospital, do patients who have a diagnosis of osteoporosis and are admitted to Nebraska Medicine hospital experience fewer injuries from falls when using a hip protector alone (no medication) compared to those who are prescribed a bisphosphonate alone (no hip protector)?"

**Identifying a Population or Patients of Interest**

When identifying the population of interest, researchers should consider both inclusion and exclusion criteria, including both demographic and medical characteristics. Demographic characteristics include traits such as age, gender, race/ethnicity, socioeconomic status, primary language spoken, and religion. Medical characteristics can include vital signs, specific diagnoses or comorbidities, or they can be medical events, such as a procedure, a visit to the emergency department, or death. Whenever possible, age, gender, and race/ethnicity should be collected in order to describe the sample in any future publications. Investigators should determine which other factors, such as pregnancy, smoking or substance use/abuse would impact a patient's inclusion in the study. Additionally, they should consider biases inherent with patients in this type of study. For example, by its nature, EHR research includes information about patients who have chosen to seek care. This population may be fundamentally different from people who do not seek care. For example, one researcher found that sicker patients had more complete data in the EHR than healthy ones.(Weiskopf, et al., 2013)
If examining differences between cases and controls, care should be taken to select a comparison group that is phenotypically similar to the cases in the population of interest. This process begins by identifying confounding variables and covariates that are thought to be related to the study outcomes. A technique called "propensity score" analysis, first developed in the early 1980's, can be used to "account for the confounding effect of these covariates and establish an association between the exposure and the outcome of interest."\textsuperscript{3,29,30} The comparison group should be similar to the cases on these covariates. Researchers should be careful not to identify too many variables on which to match cases and comparisons, which would reduce the number of available matches, and if possible, the comparison group should be larger than the group of cases.\textsuperscript{3}

**Selecting Variables**

The identification of the patient population requires the identification of some variables of interest such as age, gender, diagnoses, etc. Researchers also need to consider the study variables of interest. Generally, research questions examine the effect of x on y, the relationship between x and y, or whether y differs across different levels of x (where x is a categorical variable such as gender, disease status or the clinic location of their primary care doctor). Therefore, in addition to the inclusion and exclusion variables identified for patient inclusion in the study, researchers should determine which other variables or data points are needed to answer their question of interest, including the x and y variables.

In a randomized controlled trial, theoretically, the effects of any confounding variables (a variable that affects the outcome variable) are distributed equally across groups as a result of the randomization process and therefore do not have an undue influence on any single group or condition. Because EHR research does not allow for randomization, the effects of confounders should be addressed, if possible using other methods. The previously mentioned propensity score analysis is one way of addressing this. There are various other statistical methods that can control
for these effects. In order for these methods to work, however, these confounding variables, along with covariates, moderating and mediating variables, need to be identified and their data points recorded.\textsuperscript{31,32}

The identification of variables of interest must then be translated into fields within the EHR that contain the desired data. This process began as the research question was formulated and the components were defined. As additional variables are identified, and as links are made to fields within the EHR, the research question may be refined.

**Data Cleaning**

Because of the challenges and limitations to EHR research, often, the data that are extracted from the EHR are complex and messy. Before data are analyzed, data-cleaning and standardization steps should be taken to ensure the best possible data quality.\textsuperscript{(Terry 2009; Weiskopf et al., 2012)} Data of varying types from varying sources may have unique quality issues that should be examined. For example laboratory test results may vary in terms of sample quality, but indication of quality is not typically included in data extraction.\textsuperscript{3} Table 1 indicates some common categories of data along with specific data quality issues to consider for each.
<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
<th>Common issues to consider</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>Age, gender, ethnicity, height, weight</td>
<td>Highly sensitive data requiring careful de-identification. Data quality in fields such as ethnicity may be poor</td>
</tr>
<tr>
<td>Laboratory</td>
<td>Creatinine, lactate, white blood cell count, microbiology results</td>
<td>Often no measure of sample quality. Methods and reagents used in tests may vary between units and across time</td>
</tr>
<tr>
<td>Radiographic</td>
<td>X-rays, computed tomography (CT) scans, echocardiograms</td>
<td>Protected health information, such as names, may be written on slides. Templates used to generate reports may influence content</td>
</tr>
<tr>
<td>Physiologic data</td>
<td>Vital signs, electrocardiography (ECG) waveforms, electroencephalography (EEG) waveforms</td>
<td>Data may be pre-processed by proprietary algorithms. Labels may be inaccurate (for example, “fingerstick glucose” measurements may be made with venous blood)</td>
</tr>
<tr>
<td>Medication</td>
<td>Prescriptions, dose, timing</td>
<td>May list medications that were ordered but not given. Time stamps may describe point of order not administration</td>
</tr>
<tr>
<td>Diagnosis and procedural codes</td>
<td>International Classification of Diseases (ICD) codes, Diagnosis Related Groups (DRG) codes, Current Procedural Terminology (CPT) codes</td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Often based on a retrospective review of notes and not intended to indicate a patient’s medical status. Subject to coder biases. Limited by suitability of codes</td>
<td></td>
</tr>
<tr>
<td>Caregiver and procedural notes</td>
<td>Admission notes, daily progress notes, discharge summaries, Operative reports</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Typographical errors. Context is important (for example, diseases may appear in discussion of family acronyms are common</td>
<td></td>
</tr>
</tbody>
</table>

From Secondary Analysis of Electronic Health Records, MIT Critical Data, 2016, Springer EBook

Kahn et al. (2012) proposed a conceptual model for examining data quality.\textsuperscript{33}

Dziadkowiec et al. (2016), expanded on the work of Kahn's team to apply this model to cleaning data extracted from an EHR, and provided sample SPSS syntax to accomplish these tasks.\textsuperscript{34}

Kahn's data concepts, along with Dziadkowiec's reconceptualization, can be seen in Table 2.
### Table 2: Techniques for Assessing Data

<table>
<thead>
<tr>
<th>Key Data</th>
<th>What to Assess</th>
<th>Assessment Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concepta</td>
<td></td>
<td>Coding and recoding checks and frequency analysis.</td>
</tr>
<tr>
<td>Attribute</td>
<td></td>
<td><strong>Example:</strong> Do the responses match the predefined coding pattern? Are there variables currently coded as string variables that can be recoded into a value-labeled numeric variable?</td>
</tr>
<tr>
<td>Domain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constraints</td>
<td></td>
<td>Missing data analysis.</td>
</tr>
<tr>
<td>Missing data</td>
<td></td>
<td><strong>Example:</strong> Are the missing values logical or is there a potential source of input or output error that should be considered?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Compare patient IDs after a merge or compare the same patient ID on demographic variables.</td>
</tr>
<tr>
<td>Relational</td>
<td></td>
<td><strong>Example:</strong> Prior to merging two data sets based on the primary key, are there missing values where you would expect them to be? Do the number of rows and variables in the merges data set add up to what were in the data sets that the merge comprised?</td>
</tr>
<tr>
<td>Integrity Rules</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between</td>
<td></td>
<td>Compare results of merging data sets (by comparing primary keys or patient IDs) between sites.</td>
</tr>
<tr>
<td>database</td>
<td></td>
<td><strong>Example:</strong> Consider adding variables that code sites prior to merging so that errors can be easily traced back to the correct premerge file.</td>
</tr>
<tr>
<td>consistency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between site</td>
<td></td>
<td></td>
</tr>
<tr>
<td>consistency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Historical Data Rules</td>
<td>Time interval coding</td>
<td>Make sure that the time intervals are coded in the same units for all records and capture the desired time frame.</td>
</tr>
<tr>
<td>-----------------------</td>
<td>----------------------</td>
<td>----------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Data Rules</td>
<td>Time stamps</td>
<td>Check that time stamps fall in expected intervals (weekly or monthly) and don’t exceed a preestablished frequency.</td>
</tr>
<tr>
<td>Event sequences per person and within a site</td>
<td>Make sure that the last event time occurs before the first event time.</td>
<td></td>
</tr>
<tr>
<td>State-Dependent Objects</td>
<td>Event sequences per person and within a site</td>
<td>Example: Verify multiple events to ensure that the primary event recording is accurate.</td>
</tr>
<tr>
<td>Event sequences per person and within a site</td>
<td>Sequence timing by event</td>
<td>Make sure that events have appropriate concurrent event times.</td>
</tr>
<tr>
<td>State-Dependent Objects</td>
<td>Event sequences per person and within a site</td>
<td>Example: Ensure that not only the event dates are correct, but for those events that occur on the same day, ensure that the recorded times make sense.</td>
</tr>
<tr>
<td>Event sequences per person and within a site</td>
<td>Check to make sure that events that depend on a previous event (treatment that follows a certain diagnosis) make sense.</td>
<td></td>
</tr>
<tr>
<td>Qualifying events</td>
<td>Example: An individual who gets admitted to psychiatric ED needs to have a psychiatric diagnosis; find both the diagnosis and psychiatric ED visit variable and compare their frequencies.</td>
<td></td>
</tr>
<tr>
<td>Dependent events</td>
<td>Example: Patients with discharge and departure events should also have arrival event information.</td>
<td></td>
</tr>
</tbody>
</table>

*Based on Kahn, et al., 2012, Adapted from Dziakowiec, et al., 2016*
CHAPTER 5: ETHICAL CONSIDERATIONS

The use of EHR data for research raises some specific ethical uses that must be considered. These records likely contain protected health information (PHI) for patients. Because many health systems require patients to opt out of the use of their data for research, rather than require active consent to opt in, patients included in these studies may or may not have explicitly consented to the use of their data. Every attempt must be made to adhere to ethical use standards for EHR research.

Organizations should have procedures in place that protect patient privacy and confidentiality while allowing for the use of EHRs to obtain representative research data. Such procedures should stipulate that the data remain inside the organization and that the extracted data are restricted to the needs of a given project. Data use agreements between researchers and the data caretakers at the organization should specify how data use, storage and de-identification are carried out, and researchers should agree to strict confidentiality restrictions. Data should be encrypted and password-protected when possible, and care should be taken that data queries are not so specific that they yield narrow results that could theoretically be re-identified. Access to extracted raw data should be restricted to only specific members of the research team.

The use of data from multiple sources provides additional ethical consideration in that patient identifiers are usually needed in order to merge the data files. Researchers should have a plan in place to strip the identifiers from the data file as soon as possible. Unstructured text fields may contain identifying information, so deidentification efforts should incorporate plans to address unstructured data as well.

Whenever possible, the use of limited data sets or deidentified data sets should be used. Limited data sets may have certain identifiers, such as names or full street addresses, stripped or masked but may still contain some PHI, such as zip codes or dates, as needed to conduct the
research. The use of limited data sets offer more patient protection than fully identified data, but not as much as fully deidentified data.11

Fully deidentified data sets have all patient identifiers (patient name, medical record numbers, insurance membership numbers or other account numbers, address, zip codes, all dates, social security numbers, phone numbers and email addresses, vehicle license plate numbers, facial images, etc.) removed from the data before a researcher has access to it. There is no discernable means to reidentify a patient. Dates are shifted by a random amount for each patient and identifying account numbers, such as medical record numbers, are replaced with a randomly generated number. Rare diseases, conditions or events are excluded.11 The U.S. Department of Health and Human Services has specific guidelines for deidentifying data in compliance with the Health Insurance Portability and Accountability Act.35

Using deidentified data is obviously the most desirable strategy for conducting EHR research; however, the creation and maintenance of deidentified data is very resource-intensive. It requires sophisticated programming expertise and time to establish a deidentified data set, and additional resources are needed to extract the raw data for researchers to analyze.19 Appendix B includes step-by-step instructions for using UNMC’s deidentified electronic health record database.

CHAPTER 6: CONCLUSION

Appendix C contains a flow chart detailing the process of using an EHR for research. The use of electronic health records (EHRs) for research has both strengths and weaknesses. It can be an efficient means to conduct healthcare research, but at the expense of researcher control. On the surface, it seems to be a quick, easy, inexpensive method to conduct healthcare research. However, data extracted from EHRs can be very unreliable. There are very likely to be flaws in the data that will impact the research interpretations. Researchers should develop a clear
understanding of the pitfalls inherent in EHR research before beginning a project. As is the case with most research, many of the drawbacks can be reduced with careful preparation, formulation of a research question, procedures and data management.
REFERENCES


APPENDIX A: Questions to Evaluate the Quality of Retrospective Data Studies (from Motheral et al., 2003)

<table>
<thead>
<tr>
<th><strong>Data Sources</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Relevance</strong></td>
<td>Have the data attributes been described in sufficient detail for decision makers to determine whether there was a good rationale for using the data source, the data source’s overall generalizability, and how the findings can be interpreted in the context of their own organization?</td>
</tr>
<tr>
<td><strong>Reliability and Validity</strong></td>
<td>Have the reliability and validity of the data been described, including any data quality checks and data cleaning procedures?</td>
</tr>
<tr>
<td><strong>Linkages</strong></td>
<td>Have the necessary linkages among data sources and/or different care sites been carried out appropriately, taking into account differences in coding and reporting across sources?</td>
</tr>
<tr>
<td><strong>Eligibility</strong></td>
<td>Have the authors described the type of data used to determine member eligibility?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Methods</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data analysis plan</strong></td>
</tr>
<tr>
<td><strong>Design selection</strong></td>
</tr>
<tr>
<td><strong>Research design limitations</strong></td>
</tr>
<tr>
<td><strong>Treatment effect</strong></td>
</tr>
<tr>
<td><strong>Sample selection</strong></td>
</tr>
<tr>
<td><strong>Eligibility</strong></td>
</tr>
<tr>
<td><strong>Censoring</strong>: were inclusion/exclusion or eligibility criteria used to address censoring and was the impact on study findings discussed?</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td><strong>Operational definitions</strong>: are case (subjects) and end point (outcomes) criteria explicitly defined using diagnosis, drug markers, procedure codes, and/or other criteria?</td>
</tr>
<tr>
<td><strong>Definition validity</strong>: have the authors provided a rationale and/or supporting literature for the definitions and criteria used and were sensitivity analyses performed for definitions or criteria that are controversial, uncertain, or novel?</td>
</tr>
<tr>
<td><strong>Timing of outcome</strong>: is there a clear temporal (sequential) relationship between the exposure and outcome?</td>
</tr>
<tr>
<td><strong>Event capture</strong>: are the data, as collected, able to identify the intervention and outcomes if they actually occurred?</td>
</tr>
<tr>
<td><strong>Disease history</strong>: is there a link between the natural history of the disease being studied and the time period for analysis?</td>
</tr>
<tr>
<td><strong>Resource valuation</strong>: for studies that examine costs, have the authors defined and measured an exhaustive list of resources affected by the intervention given the perspective of the study and have resource prices been adjusted to yield a consistent valuation that reflects the opportunity cost of the resource?</td>
</tr>
<tr>
<td><strong>Control variables</strong>: if the goal of the study is to examine treatment effects, what methods have been used to control for other variables that may affect the outcome of interest?</td>
</tr>
<tr>
<td><strong>Statistical model</strong>: have the authors explained the rationale for the model/statistical method used?</td>
</tr>
<tr>
<td><strong>Influential cases</strong>: have the authors examined the sensitivity of the results to influential cases?</td>
</tr>
<tr>
<td><strong>Relevant variables</strong>: have the authors identified all variables hypothesized to influence the outcome of interest and included all available variables in their model?</td>
</tr>
<tr>
<td><strong>Testing statistical assumptions:</strong></td>
</tr>
<tr>
<td>-----------------------------------</td>
</tr>
<tr>
<td><strong>Multiple tests:</strong></td>
</tr>
<tr>
<td><strong>Model prediction:</strong></td>
</tr>
<tr>
<td><strong>Discussion / Conclusion</strong></td>
</tr>
<tr>
<td><strong>Theoretical Basis:</strong></td>
</tr>
<tr>
<td><strong>Practical versus Statistical Significance:</strong></td>
</tr>
<tr>
<td><strong>Generalizability:</strong></td>
</tr>
</tbody>
</table>
Using the Deidentified UNMC i2b2 Database for Research

University of Nebraska Medical Center
Omaha NE 68198
Updated 8/3/18 by
Jenenne Geske, PhD
Applying for Access

Procedures TBD

Logging On

- Once access is granted go to database link
- Login using your UNMC credentials

The i2b2 Home Screen

- The screen has 5 main sections:

  1. Navigate Terms / Find Terms: Locate search terms to query; click on folder to expand
  2. Workplace: Save queries. Contains SHARED Folder and Personal Folder. Queries in SHARED folder can be seen by any user (useful for sharing queries with collaborators, or finding generic queries that might be useful).
  3. Previous Queries: Stores all queries that you have run.
  4. Query Tool: Search engine; Defaults to 3 Groups – can add additional Groups if needed.
  5. Query Status: Provides query results

- You can enlarge the workspace for any of the sections by clicking on the "Resize Workspace" icon in the upper right hand of each section.
Finding Search Terms

- Once search criteria (variables) are identified, find the appropriate terms in Section 1. They can be located by expanding the hierarchical subfolders in the "Navigate Terms" tab, or by clicking on the "Find Terms" tab and searching for the desired term.
- The "Find Terms" tab has subtabs that allow you to "Search by Names" or "Search by Codes."
- The "Search by Names" subtab further allows you to restrict the search using the restrictors "Containing," "Exact," "Starting With," or "Ending With." It also provides options to restrict the search by category using a drop down list. Enter the search terms in the empty box, select any desired restrictions, and then click on "Find."

- The "Search by Codes" subtab allows you to search using one of a variety of coding systems, including ICD9CM, ICD10CM, SNOMED, and RxNorm.
Setting Up a Query

- Your search for terms/criteria will result in a list of possible query terms.
- Drag each of your desired terms to the Query Tool section.
- You can put multiple terms within each Group and you can use any number of groups.

Deleting a Term from the Query Tool

- To delete a term from the Query Tool, right click on the term and select "Delete."

Running a Query

- Once the terms are entered into the Query Tool, click on "Run Query"

Enter a descriptive name for your query. The default setting for query result type is "Number of patients", and since that is what the research question is looking for, this is the appropriate setting for this query. Click on "OK."
Query Results

- The result will be in the "Query Status" section, which shows that there are 623 patients who meet all three of the query criteria:

![Query Status](image)

Logical conditions ("Or" / "And")

- If there are multiple terms in the same group, the system will return the patients who meet at least one of those criteria (they must meet term/criterion "a" OR term/criterion "b" OR term/criterion "c".)
- Across groups, all of the criteria must be met for a patient to be included (they must meet term/criterion (Group 1) AND term/criterion (Group 2) AND term/criterion (Group 3).
**EXAMPLE 1: "And" condition**

*How many patients are there between 35-44 years old when they were seen, who suffered any fracture of the radius or ulna, and who reside in Nebraska?*

- All of the terms/criteria must be met in order to be a patient of interest, so each search term/criterion would be entered into a separate Group. The query would look like this:

![Query Diagram]

- The result shows that there are 623 patients who meet all three of the query criteria:

![Query Status]

*Finished Query: "35-44-813 F-Nebra@15:10:26"*
*Compute Time: 4 secs*
*Number of patients for "35-44-813 F-Nebra@15:10:26" patient_count 623+3*
EXAMPLE 2: "Or" condition

- What if the question changes to:

How many patients are there between 35-44 years old when they were seen or who suffered any fracture of the radius or ulna or who reside in Nebraska?

- In this situation, any patient who meets at least one of the terms/criteria is a patient of interest. All of the search terms are placed in the same Group:

Because patients only have to meet one of the terms/criteria, rather than all 3, the number of patients is much higher than in EXAMPLE 1. There are 1,634,352 patients who meet at least one of these criteria.
EXAMPLE 3: Combining "And" and "Or" conditions

- If the question becomes instead:

  How many patients are there between 35-44 years old when they were seen and who suffered any fracture of the radius or ulna, and who reside in Nebraska or Iowa?

- A new criterion is added; patients who live in either Nebraska OR Iowa are of interest.
- The query then changes with the addition of an "OR" term; "Iowa" is added to Group 3 to represent these additional patients of interest.

- Now, with the inclusion of the Iowa residents, the number of patients of interest increases to 684.
EXAMPLE 4: Adding additional Group boxes / Adding exclusions

- What if people who lived in areas with a zip code beginning with "630" should be excluded from the search?

*How many patients are there between 35-44 years old when they were seen, (and) who suffered any fracture of the radius or ulna, and who reside in Nebraska or Iowa, (and) excluding those who live in a 630** zip code?*

- We now have four search terms, so we need to add another "Group" box. In the bottom right hand corner of the Query Tool, click on the "New Group" button to add "Group 4." The arrows will allow you to move between the groups on your screen.

- To exclude certain categories of patients, move the exclusion term to a Group box and click on "Exclude" in the upper right hand corner of the box. A red box will appear at the bottom of the window indicating that none of the patients who meet that criterion should be included. (The first three Groups will be identical to those in EXAMPLE 3.)
EXAMPLE 5: Modifying the query result types

- What if a breakdown by gender is needed?

*How many patients are there between 35-44 years old when they were seen, who suffered any fracture of the radius or ulna, and who reside in Nebraska or Iowa, broken down by gender?*

- Click "Gender patient breakdown" in the window that pops up after clicking "Run Query," then click "OK."

- Scroll down to see that there are 396 Females and 284 Males in this group.
Additional Constraints

- The "Temporal Constraint" box in the Query Tool allows you to determine which events you would like to include in the query.
- It defaults to the "Treat all groups independently" option, which includes all qualifying events over the course of a patient's life.
- The "Selected groups occur in the same financial encounter" option restricts query results to events that occur at the same time. For example, this option would allow you to select those who are hospital inpatients and have a diagnosis of diabetes at the time of their hospital stay. This option would exclude patients who may have been diagnosed with diabetes sometime after their hospital stay because that diagnosis did not exist for them at the same time as their hospitalization.
- The "Define sequence of events" option allows you to specify the order of interest for specific events.
- In each of the Group boxes, specific date ranges of interest can also be specified by clicking on the "Dates" button.
- You can also specify a minimum number of times an event occurs before that patient will be included, by clicking on the "Occurs>0x" button. For example, a researcher may only wish to include patients who have more than 1 fractures over the past 5 years.

Add as information becomes available….

Obtaining authorized access

How to obtain raw data

Expanded Hierarchy (Appendix)
APPENDIX C: The Medical Records Research Process flow chart

**Topic**
- Identify area of interest for research
- Frame Research Question, based on PICOT format, if appropriate.

**PICOT**
- **P**: Patient, Problem or Population of interest
- **I**: Intervention
- **C**: Comparison or Control group
- **O**: Outcome
- **T**: Time Frame

**Literature Review**
- Become familiar with previous work in the area
- Identify gaps in literature
- Contact McGooian librarians for help locating relevant articles

**Research Question**
- Revise Research Question if needed based on literature review
- Novel
- Specific (include specific variables identified in previous step)
- Concise
- Arguable

**Variables**
- Identify: Condition/disease/ trait/ clinical event of interest
- Population/ Sample/ Cohort
- Control/ Comparison group (if applicable)
- Covariates/ Control Variables
- Possible sources of error in data
- Ways to reduce error
- Needed demographics
- Inclusion / Exclusion Criteria
- Search Parameters
  - Period of time
  - Specific clinic/ department
  - Is first/ last/ average observation

**Search Terms**
- For each of the variables identified in previous steps, determine the type of:
  - health care activity
  - event
  - data value
  - that will be the source of information (e.g. specific test values, diagnosis code (e.g. ICD-9 or ICD-10, medication, referral)

**Data Source(s)**
- Can data be obtained from deidentified i2b2 database, or are identifiers needed?
- Are data needed from other sources besides the EHR
- Are any prospective data being collected?
- If identifiers are needed, or if prospective data is needed, submit IRB application (not necessary for clinical QI projects)

**Obtain Data**
- Assess needed sample size
- Obtain i2b2 access
- Perform query
- Send MRNs to data people for raw data
- After IRB is approved, submit data request form ([https://www.unmc.edu/cctr/resources/ehr/index.html](https://www.unmc.edu/cctr/resources/ehr/index.html))

**Analyze Results**
- Enlist help from statistician if necessary

**Share Results**
- Present at conference
- Submit manuscript for publication