OREGON HEALTH & SCIENCE UNIVERSITY
SCHOOL OF MEDICINE – GRADUATE STUDIES

Capstone Project

Finding Healthy Patients from Electronic Health Records for Research Control Subjects in Clinical Trial Recruitment

Master of Biomedical Informatics
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
Portland, Oregon

Term: Fall 2016
Student: Katharine Fultz Hollis, MS
Capstone Advisor: Tim Burdick, MD, MSc

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's Capstone Project of

Katharine Fultz Hollis

Finding Healthy Patients from Electronic Health Records for Research Control Subjects in Clinical Trial Recruitment

has been approved

_____

Tim Burdick, MD, MSc

Introduction

Using electronic health records (EHRs) to find patients for clinical trials is both exciting and challenging. “EHRs often do not tell a complete patient story” (1) for a variety of reasons, and it has been very difficult to find types of patients such as those you might want as control patients (primarily patients with few health problems and no serious chronic diseases). However, EHR data can be used to find cohorts of patients who suffer from specific diseases and to identify patients with certain problems for clinical trials. Essential to clinical trial recruitment is the identification of healthy patients (or “control” patients) for research studies and these cohorts are usually difficult to find in a health care EHR system. But EHR data can be searched using an algorithm of ICD-10 codes, some medications, and elimination of certain kinds of patients and it becomes possible to identify healthy participants for clinical trials. This capstone project surveyed the literature on computer algorithms to find control patients for clinical trial recruitment, analyzed the data retrieved from the run of one algorithm to find “healthy” patients in the Oregon Health & Science University (OHSU) EHR system, and showed similar searches for healthy patients using OHSU’s Cohort Discovery Tool and Epic Real-time Reporting Workbench queries. Whether a researcher can find enough data from the EHR to find patients that could be described as having a healthy patient phenotype is complicated and difficult to assess now given the lack of data available in an EHR. However, preliminary results from the healthy patient algorithm suggest that the EHR can retrieve patients as possible candidates for a clinical trial patient registry and developing algorithms for patients lacking chronic diseases is possible.

Background

OHSU is a quaternary care, academic medical system located in Portland, Oregon. OHSU has an adult and a pediatric hospital, 28 locations throughout Oregon including a 100-acre campus on Marquam Hill in Portland, 1,017,964 patient visits in 2015, five graduate schools, and $376 million in research dollars granted in 2015 (2). The Epic© EHR was implemented at OHSU in 2004 in the ambulatory setting and has been expanded to all clinical settings (over 100 clinics in the system). According to the OHSU Epic help desk, the patient portal, tethered to the EHR, was launched and implemented in 2005 and 2006.

OHSU's Oregon Clinical & Translational Research Institute (OCTRI) has the Clinical and Translational Research Center (CTRC) that maintains an institutional review board (IRB) approved registry of participants interested in participating in OHSU medical research. The registry contains basic demographic information and limited medical history to allow researchers to contact potentially eligible research participants. The registry also includes a biorepository with serum, plasma, urine and saliva from healthy subjects. Samples can be made available for researchers who need control samples for their research projects (3).

Recruitment for potential participants for the CTRC's registry has been done by phone, mail, and OHSU sponsored recruitment events. OHSU Principal investigator Mary Samuels, MD presented in her 2014 Institutional Review Board proposal for the registry the difficulty in recruiting clinical trial participants and recruiting healthy participants for control subjects. She wrote that an analysis (2006-2009 data, 374 studies) showed that 31.1% of OHSU clinical research studies enrolled zero or one subject before being terminated. OHSU wastes at least $1 million per year on these studies (4). Repeat

analysis using 2010-2012 data showed little progress, with 39% of 447 studies still enrolling 0-1 subject. This is a significant underestimate of under-enrollment, since there are further drains due to studies that enroll more than one subject, but fewer than needed for optimal results. The underlying cause of under-enrollment is multi-factorial, but includes inadequate time, experience and resources to recruit subjects effectively. Electronic health record (EHR) data can be retrieved to find cohorts of patients who suffer from chronic and other diseases and the data has been valuable for retrieving groups of patients. Cohorts of healthy patients as control patients are difficult to find in a health care system because by the fact that they are healthy, they might not regularly appear at the hospital or clinic. There might be less information on the patient's record that would qualify that patient to be healthy. However, because clinical trials often do not have enough subjects (4), it is important to try and use the now increased availability of EHR data to find healthy participants for possible clinical trials.

Materials and Methods

Our study included three activities:

1) a detailed review of past studies and research on clinical trial algorithms including what might be defined as a "healthy" patient for control studies;
2) the run of a computer algorithm on November 3, 2015 to find healthy patients; and
3) attempts to replicate a similar to 2) search for healthy patients using OCTRI Cohort Discovery and the OSHU Epic Reporting Workbench.

In our study, we wanted to first determine if there had been studies on the healthy phenotypes in the scientific literature. Very little information exists on recruiting control patients in general and very difficult to find studies that focus on identifying healthy patients using computers (Luzurier et al touch upon strategies for recruitment of healthy

volunteers but not identification (5)). Little information exists in the medical literature about using algorithms to capture relevant data on healthy patients from EHRs and an extensive literature search with PubMed, Cochrane, and Scopus was conducted from June 2016 to September 2016. Besides the lack of studies about EHR recruitment and especially "healthy" patient recruitment, EHR data does not contain all the data you might need to define a certain kind of patient who has few medical problems. Also, most healthy people tend not to come to the clinic at all except for annual checkups and immunizations. However, if we were looking for a disease cohort, there have been studies on algorithms that help identify patients with a disease. A 2012 study by Beauharnais et al has shown that "using a computer algorithm to identify eligible patients for a clinical trial in the inpatient setting increased the number of patients screened and enrolled, decreased the time required to enroll them, and was less expensive" (6). This reference outlined steps to identify a cohort from an in-patient population.

The literature searches included PubMed, Scopus, Cochrane Review, and Google Scholar. Most articles review computable disease phenotypes from patient data but there was very little about using computer algorithms to find healthy patients or control subjects.

Here is a selection of studies that we found most useful for learning about computer algorithms for clinical trial recruitment:

a) Köpcke F, Prokosch H-U. Employing computers for the recruitment into clinical trials: a comprehensive systematic review. 2014.

   Köpcke presented a very detailed systematic review of computer algorithms for clinical trials and his study was one of the few that focused on clinical trial recruitment. His systematic review is one of the few available on the topic of algorithms and clinical recruitment methodology. Köpcke noted a systematic review by Cuggia in 2009 that

concluded it was still difficult "to make any strong statements about how effective automatic recruitment is, or about what makes a good decision support system for clinical trial recruitment (7)." Köpcke found that in the final pool of 101 relevant articles found on the subject of clinical trial recruitment support systems (CTRSS) "most articles describe the characteristics and operating principles of their CTRSS reasonably well, but all lacked in some regard. Intermediary criteria representation, terminologies of the patient data, and an evaluation of the system's effects were often missing. Many authors present prototypes of their CTRSS directly after finishing its design and fail to report on its outcome and usage."

b) Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. 2013.

The Newton study provided details on using the electronic medical record (EMR) to find genomic data and to validate the phenotype. The authors include a very informative table of positive predictive value for phenotype case and control algorithms. One conclusion in the Newton study is important for our study: "EMRs cannot capture all nuances of patient–provider interactions, but they are extremely useful resources for well designed, informative clinical studies. Accurate EMR capture of diagnosis, laboratory, and medication data, supplemented with text-mining tools and NLP (natural language processing), can provide excellent phenotype data for genomic studies, including GWAS. However, even with advances and new approaches, the heterogeneity in EMRs means that phenotype validation will remain an important aspect of their use (8)."

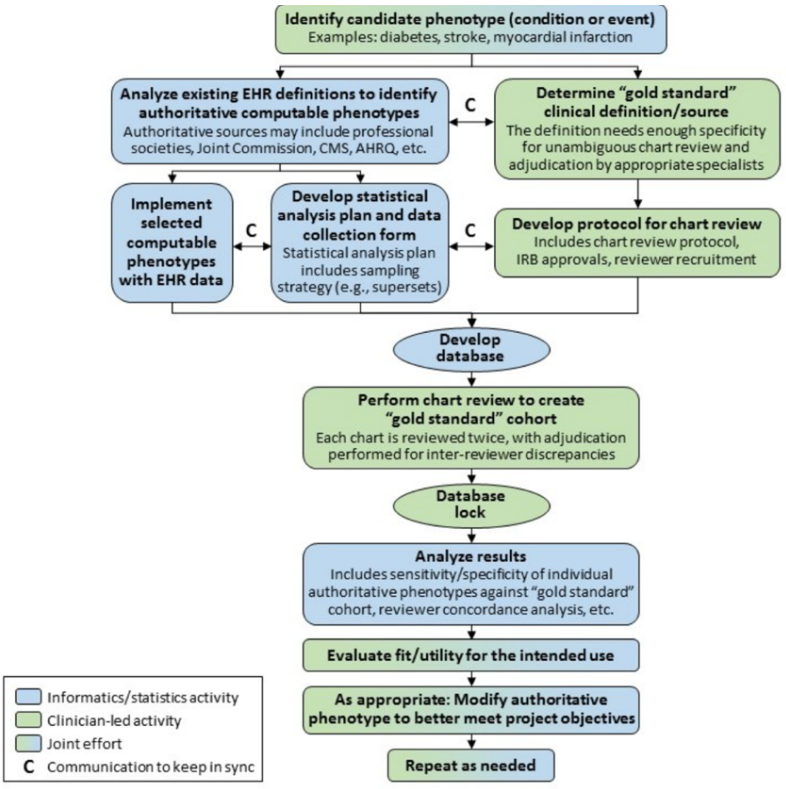
c) Rasmussen LV, Thompson WK, Pacheco JA, Kho AN, Carrell DS, Pathak J, et al. Design patterns for the development of electronic health record-driven phenotype extraction algorithms. 2014.

Like Newton, Rasmussen focused his study on phenotype algorithms using data from the electronic Medical Records and Genomics network

(eMERGE) (9). Phenotypes created by the eMERGE network are publicly available on the Phenotype KnowledgeBase website (PheKB, http://www.phekb.org). Interestingly, there were algorithms for simple attributes like height and weight but variation in records made those as difficult to validate as those algorithms for diseases. PheKB, by the way, does not have a healthy patient phenotype.

d) Richesson R. Electronic Health Records-Based Phenotyping | Rethinking Clinical Trials®. 2015.

Richesson's book (10) provided a helpful graphic for the phenotype development shown here:



Phenotype evaluation process. AHRQ, Agency for Healthcare Research and Quality;

Figure 1. Richesson phenotype evolution process.

The phenotype evaluation process for a healthy patient phenotype would require many reviews of the healthy patient charts to see common characteristics. Richesson attempts to show that gold standards can be

> developed for the accuracy of the phenotype (through modification to find as close to accurate definition of the disease). A "healthy patient" phenotype might be difficult to describe as existing EHR definitions of "healthy" might not be available. A search process for a healthy patient in the EHR might include eliminating many from the total patient population as you might define a patient as being NOT "healthy."

Perhaps the most difficult part of the phenotype development is the determination of a gold standard clinical definition for "healthy" but difficulty should not mean impossibility (yet).

The algorithm to find a healthy patient phenotype in the Research Data Warehouse (containing EHR data from the OHSU system) at OCTRI was developed by Tim Burdick, Mary Samuels and Peter Beninato, and the description for the algorithm is available in **Appendix 1** (henceforth known as the Burdick/Samuels/Beninato algorithm). To find a healthy patient in a defined set of OHSU EHRs, we were limited to certain departments that allowed access to de-identified patient data and removal of opt-out patients. Most of the ICD-10 codes we used were to exclude subjects, except we also allowed some sub-codes of minor conditions within some of the broader codes. We focused on patients who were listed with the following acceptable medications: over-the-counter drugs, antibiotics, hormonal or other contraception, allergy, non-steroidal anti-inflammatory drugs, and acetaminophen. Finally, there were groupings of patients that were acceptable for inclusion and these groups in turn had to have some excluded because of medications and diagnosis. It is not an easy program to set up and the SQL instructions include directions to include certain patient encounter types and exclude provider credentials that are from a certain department.

A total of 659 parent ICD-10 codes were acceptable for inclusion in the algorithm to search for possible healthy patients (see **Appendix 2** for the complete ranges of codes). Using SQL, Burdick, Samuels and Beninato retrieved 858 records of possible healthy patients from a query to a Clarity database on November 3, 2015.

The final activity was to search for patients in the OHSU Epic system that might be classified as "healthy" and use some of the parameters established by the team when they searched on the Clarity database. Could we replicate the system to eliminate certain ICD-10 codes and medications and then from that group include those who have few problems and medications that healthy people might use (such as ibuprofen)? First, we tried the OCTRI's Cohort Discover Tool (CDT). The CDT is "a web-based tool for finding sets of interesting patients in OHSU's Epic Electronic Health Record (EHR) for preparatory to research purposes (11)."

Using CDT, we attempted to replicate the query we developed with the algorithm and tried to find a way to search for the variables we outlined in **Appendix 1**. CDT Cohort Discovery utilizes open-source software funded by an NIH cooperative agreement with Partners HealthCare System through the National Center for Biomedical Computing (NCBC) program called i2b2 (Informatics for Integrating Biology and the Bedside). We could use only ICD-10 codes in our queries and the medication categorization in CDT uses OHSU's formulary, which has many more detailed top level categories (e.g., Cephalosporin – 1st Generation) than the Epic therapeutic classes used by the SQL query on November 3, 2015. We were also limited with the CDT because the tool did not allow us to exclude some departments that did not allow access to de-identified patient data from certain primary care providers.

Another search option we tried was Epic's Reporting Workbench (RW), a program that is used to report on small volumes of real-time and near-real-time data about patients (the healthy patient algorithm was run on the Clarity database that is used to report on large volumes of historical data). One disadvantage to RW is that the pool of patients is small due to the real-time nature of the data set. Whether RW contains patients that did not opt-out for possible research studies is also hard to determine. We also wanted to try searching clinical notes for instances of "healthy" or variations of healthy like "excellent health." However, Rosenbloom et al have described in detail the difficult of searching clinical documentation in general and in the EHR: "structured entry systems typically have not enjoyed long-term or widespread adoption. McDonald and Ash have demonstrated that structured entry adoption (of clinical notes) may be hampered by user interface complexity, inflexibility for documenting unforeseen findings, lack of integration with other clinical applications, and deficiencies in the underlying data model (12)."

Searches for "healthy" patients were performed on November 21 through November 25, 2016. We were unable to exclude certain patients using the exclusion criteria outlined in **Appendix 1** because the queries on RW do not appear to allow for exclusion (like using the Boolean term NOT to exclude certain diagnosis codes or ICD-10 codes). Yet we wanted to see if RW could be an easier way to search (easier than creating a program to search the OHSU EHR) and use the simple search interface created by Epic. Several searches were conducted on the interface called "RSCH Find Patients – Research" shown in **Figure 2**:

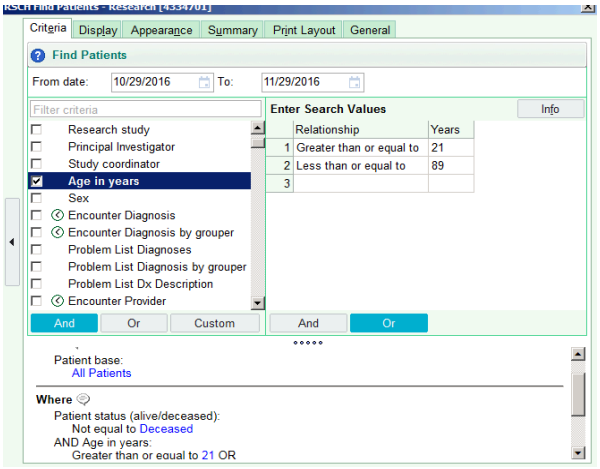


Figure 2. A screenshot of OHSU Research Workbench query

We attempted six (6) searches during the week of November 21, 2016 and we did several variations of the following search to see how many patients would be retrieved (**Figure 3**). We did a search for a month of data from the EHR in **Figure 3** and then repeated the search from 6 months of data (6/1/2016 to 11/30/2016) and the search retrieved the same number of patients (and the same patient medical record numbers).

HTML Summary Find Patients between 10/29/2016 and 11/29/2016
From
Patient base:All Patients
Where
Patient status (alive/deceased):
Not equal to Deceased
AND Age in years:
Greater than or equal to 21 AND
Less than or equal to 89
AND History: Medical:
Healthy adult on routine physical examination
And where
Authorized service areas:
Oregon Health & Science Univ [2]

Figure 3. OHSU Research Workbench search

Results

Literature Review

The literature search for defining the "healthy patient" did not retrieve much information about clinical trials attempting to find control patients using algorithms. Köpcke (7) developed a very interesting PubMed search that I modified on May 30, 2016 to search for healthy patient cohorts:

("clinical trial"[All Fields] OR "clinical trials"[All Fields]) AND ("eligibility"[All Fields] OR "identification"[All Fields] OR "recruitment"[All Fields] OR "accrual"[All Fields] OR "enrollment"[All Fields] OR "enrolment"[All Fields] OR "screening"[All fields]) AND (("participants"[All Fields] OR "cohort"[All fields]) AND "healthy"[All Fields]) AND ("electronic"[All fields] OR "computer"[All fields] OR "software"[All fields] OR "automatic"[All Fields]) AND ("2011/06/03"[PDAT] : "2016/05/31"[PDAT])

The results from the above search did not find any article that focused on clinical trial methodology for finding control patients. The following table shows the results of variations of the search for "healthy patient" phenotypes or searches for recruitment of control patients using computers. Results are from PubMed and—with some modification—Scopus, Cochrane, and Google Scholar are in **Table 2.**

Date	Search String	Results
8/17/2016	(("Electronic Health Records"[Majr]) AND "Phenotype"[Majr]) AND "Clinical Trials as Topic"[Majr]	0 references
8/17/2016	"Electronic Health Records"[Majr] AND "Patient Selection"[Majr] AND "Clinical Trials as Topic"[Majr]	22 references but none specifically about algorithms or control patients.
9/16/2016	(("Electronic Health Records"[Majr]) AND "Phenotype"[Mesh]) AND "Case-Control Studies"[Majr]	0 references
9/16/2016	("Healthy Volunteers"[Majr]) AND "Algorithms"[Majr]	0 references
9/16/2016	clinical trial recruitment AND "Algorithms"[Majr]	31 references but most are very technical about predictive modeling and we were not looking for this type of research.
9/16/2016	("Controlled Clinical Trials as Topic"[Majr]) AND "Algorithms"[Majr]	80 references but most about disease cohorts and nothing about control patients.
9/16/2016	("Controlled Clinical Trials as Topic"[Majr]) AND "Algorithms"[Majr] AND recruitment	4 references and found Beauhamais et al that discussed more of the cost-effectiveness of searching by algorithm.

Table 2. Queries and results for healthy patient recruitment.

Healthy Patient Algorithm

The Burdick/Samuels/Beninato algorithm, developed at OHSU to find healthy patients to recruit for a clinical trial registry, was successful in 2016 to identify healthy patients to send MyChart messages for recruitment to the regisry. 482 subjects were randomly chosen from the 858 records retrieved by running the algorithm on the Clarity database comprised of OHSU EHR data. A manual chart review was conducted with every 10 of 482 patients (48 charts) and it was concluded that only one patient in 48 charts should have been excluded as the diagnosis of epilepsy was in the free text of the provider's note

and not in the problem or medication lists. We could not determine the false negatives of this population and then determine what is the sensitivity of the algorithm (probability that the SQL query chose healthy patients in the collection).

While the purpose of our study here was to show the value of the algorithm in retrieving healthy patients from the OCTRI Research Data Warehouse, there is an additional study of the cost-effectiveness of using the Burdick/Samuels/Beninato algorithm to compare with other methods and that study will be published soon. The study shows the cost-effectiveness of recruitment via patient portal (using the algorithm to find healthy subjects and send MyChart recruitment letters), mail, and phone in a randomized sample of patients who all had active patient portal accounts.

OHSU Research Workbench Results

There were 6 variations of the search done in Epic's Research Workbench (RW) at OHSU and only two searches provided 39 patients with the search outlined in **Figure 3** (the only difference in the two searches was one searched with a date range of one month and the other with a date range of 6 months). All the patients had the variable "History: Medical: Healthy adult on routine physical examination" and most of these patients in a chart review could be classified as "healthy" except for perhaps a few that might have less serious chronic conditions. Unfortunately, the search term "tobacco use equals never" could not be used in this set as including that variable eliminates all the records even those with tobacco use being negative. So, retrieving all those patients from the search in **Figure 3**, we listed the tobacco use column for those patients and choose those that were listed as "never" or "history of smoking." The final list of 31 patients appears in **Appendix 3**. The search on RW could not remove certain providers as could be done in the Burdick/Samuels/Beninato algorithm. The best search using ICD-10 codes had to be

done using only one variable and that was Z00.00, Encounter for general adult medical examination without abnormal findings (OHSU Epic defines this as "healthy adult on routine physical examination"). Also, once more variables were added to RW, the searches became very slow and did not retrieve patients. The best way to search was to include only two terms at a time.

None of patients listed in **Appendix 3** had serious chronic diseases (but one had a long problem list) and some of them had patient portals (MyChart on OHSU). The following list here shows the number of healthy patients found after chart review of the 39 patients, and the results of the chart review:

Number of "Healthy" Patients found in Query: 39
Number of Patients "Not Healthy" based on chart review of 39 patients found by query: 8
The total number of patients in the EHR dataset of the query: 10066
MyChart activated: 11; MyChart declined: 3; MyChart pending: 16.

In comparison to the algorithm on Clarity, the RW search provided fewer healthy patients with MyChart activated but the two variables of age and the Medical: History value of healthy adult on routine examination produced what could be called a cohort of healthy adults. In general, for the review of the 39 charts, only a few patients had problems listed and the most serious problems included medullary thyroid cancer (1 patient) and tobacco use (7 patients). The following **Table 3** lists most of the terms appearing in the 31 charts (after chart review) of patients found with the RW query in **Figure 3.**

EHR Variable	Clinical Term or Name	Number of Patients
Allergies	**Sulfonamide Antibiotics**	**1**
	Doxycycline	**1**
	Penicillin	**2**
	Tetanus Vaccines and Toxoid	**1**
	Azithromycin	**1**
	Tylenol (Benzocaine]	**1**
	Ibuprofen	**1**

EHR Variable	Clinical Term or Name	Number of Patients
Problem List	**Propionibacterium infection**	**1**
	Dental Problems	**2**
	Cysts	**2**
	Fractures	**3**
	Urinary Tract Infection	**2**
	Tear of PCL	**1**
Medications	**adapalene-benzoyl peroxide (EPIDUO) 0.1-2.5 % topical gel with pump**	**1**
	acetaminophen 500 mg oral tablet	**4**
	lidocaine 4 % mucous membrane solution	**1**
	trimethoprim-sulfamethoxazole (BACTRIM) 80-400 mg oral tablet	**2**
Surgical History	**Wisdom Tooth Extraction**	**1**
	Surgery on ear tubes as child	**1**
	Adenoidectomy	**1**
	Jaw surgery	**1**

Table 3. Terms found in 31 "healthy" patient charts.

Overall, neither the Problem List or the Surgical History of the 31 patients contained information that would indicate that the patient had severe medical problems. The question might be well maybe severe problems were not recorded or what has been described well in the medical literature that ICD-10 codes are not sufficient and patient history data is also insufficient (13). In some instances, to search for problems patients might have is easier but when the search is for patients with no problems it is harder to identify the exact attributes of the no problem patient.

Discussion

To find "healthy" patients to recruit for clinical trials, we found that using both a complicated SQL algorithm (Burdick/Samuels/Beninato algorithm) and trying a variation on Epic's RW was successful to find what might be possible control study subjects to recruit. This study did not necessarily define the "healthy" patient phenotype but we did learn that developing good ways to capture the data from either the data warehouse or the

OHSU Epic EHR system was possible and not an easy task or a task easily to replicate (several runs need to be done including a run on a non-OHSU set of patient records). A researcher would have to go through several searches on RW to determine if for the most part "Healthy adult on routine physical examination" as a value in "History: Medical" would retrieve healthy patients 80% of the time (if our one time RW search is statistically valid). Due to not being able to determine false negatives (i.e., those records deemed not healthy by the query but were identified as healthy in a chart review) of the RW search or the Burdick/Samuels/Beninato algorithm search, we cannot define the probability that the query will identify healthy patients among those in the hospital database.

As Hripcsak and Albers noted in 2013, "as we move to large scale mining of the EHR, defining the queries has become a bottleneck" and the search for an accurate cohort for a study from using a computer algorithm can take a long time (14). They mention that the process of accurately defining a phenotype can take years especially when there are challenges like completeness and accuracy of the EHR data. In the case of finding control subjects from EHR data, there is difficulty defining a healthy person. The algorithm on Epic Clarity and the Epic RW can retrieve patients that have fewer diseases (and none chronic at least in regards to what appears in the EHR) and fewer appointments for medical problems.

A patient who has been classified as a healthy adult in a routine examination could be a candidate for a control group for a research study. But even a simple search on medical history does not easily find a prospect for the control group in a clinical trial. The best method for finding a healthy patient requires detailed steps to retrieve patients who don't have chronic disease, never have smoked, be of a certain age, and in the case of the Burdick/Samuels/Beninato algorithm, be treated by specific providers. It would be

important to continue the phenotype development process as noted by Richesson as over time more patients could be identified as healthy and be candidates for the control groups. When patients start using MyChart at OHSU more there will hopefully be more success to recruit patients via a patient portal.

Conclusion

While a survey of the medical literature found some very significant studies on using algorithms for clinical trial recruitment, no studies could be found on detailed algorithms to find healthy patients for control subjects in clinical trials. The Burdick/Samuels/Beninato algorithm at OHSU, however, provides a good framework to develop a healthy patient phenotype and additional work on this algorithm shows some success in recruiting patients for control groups in clinical trials. While Epic's RW queries cannot as accurately duplicate the complicated queries on Clarity, more work should be done on this method to query the EHR for research cohorts and develop a simple way to search for healthy patients for control groups in clinical trials.

References

1. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PRO, Bernstam E V, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013 Aug;51(8):30–7.
2. OHSU Who We Are [Internet]. OHSU Health. 2016 [cited 2016 Nov 2]. Available from: http://www.ohsu.edu/xd/health/who-we-are/index.cfm
3. OCTRI Research Volunteer Registry [Internet]. OHSU OCTRI Homepage. 2016 [cited 2016 Nov 2]. Available from: http://www.ohsu.edu/xd/research/centers-institutes/octri/resources/octri-research-services/research-volunteer-registry.cfm
4. Kitterman DR, Cheng SK, Dilts DM, Orwoll ES. The prevalence and economic impact of low-enrolling clinical studies at an academic medical center. Acad Med. 2011 Nov;86(11):1360–6.
5. Luzurier Q, Damm C, Lion F, Daniel C, Pellerin L, Tavolacci M-P. Strategy for recruitment and factors associated with motivation and satisfaction in a randomized trial with 210 healthy volunteers without financial compensation. BMC Med Res Methodol. 2015 Jan 5;15:2.
6. Beauharnais CC, Larkin ME, Zai AH, Boykin EC, Luttrell J, Wexler DJ. Efficacy and cost-effectiveness of an automated screening algorithm in an inpatient clinical trial. Clin Trials. 2012 Apr;9(2):198–203.
7. Köpcke F, Prokosch H-U. Employing computers for the recruitment into clinical trials: a comprehensive systematic review. J Med Internet Res. Journal of Medical Internet Research; 2014 Jan;16(7):e161.
8. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc 2013.
9. Rasmussen L V, Thompson WK, Pacheco JA, Kho AN, Carrell DS, Pathak J, et al. Design patterns for the development of electronic health record-driven phenotype extraction algorithms. J Biomed Inform. 2014 Oct;51:280–6.
10. Richesson R. Electronic Health Records-Based Phenotyping | Rethinking Clinical Trials®. Uhlenbrauck G, editor. Rethinking Clinical Trials®. NIH Health Care Systems Research; 2015.
11. OCTRI Oregon Health & Science University. Cohort Discovery: User's Guide version 1.6. Portland, OR; 2016.
12. Rosenbloom ST, Stead WW, Denny JC, Giuse D, Lorenzi NM, Brown SH, et al. Generating Clinical Notes for Electronic Health Record Systems. Appl Clin Inform. Germany; 2010 Jan;1(3):232–43.
13. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc. Jan;21(2):221–30.
14. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013;20(1).

Appendix 1

REQUEST FOR DATA FROM OCTRI RESEARCH DATA WAREHOUSE

Project: Samuels_Burdick_Healthy_Patient Investigator: Mary Samuels, (Tim Burdick)
Study Coordinator: n/a
IRB # 10709 – OCTRI-PRJ-3772

Specifications

Overview:
To identify Subjects whose health record does not contain any chronic/debilitating DXs. And whose medication records are for over-the-counter, or medications that are not indicative of a serious illness.

General
1. Fully identified datasets for patient recruitment purposes.
2. Proof of Concept one time pull of information from Epic
(w/possibly repeated data pulls as requested)
3. The eventual intent is to possibly recruit healthy subjects via MyChart emails.

Specifics
For each patient pull the following fields:

1. Demographics:
Fields to be included in dataset:

a. PAT_ID
b. MRN_CD
c. PAT_NAME
d. AGE
e. GENDER_NM f. NIH_ETHNCTY
g. NIH_RACE
h. PROV_NAME
i. PROV_PRIM_ADDR1
j. PROVIDER_CREDENTIAL_DEPT
k. SPECIALTY
l. PRIMARY_DEP_NAME
m. ACTIVE_STATUS
n. PROV_TYPE
o. FM_PROVIDER_LIST_FLG
p. ADDR_LN1 q. ADDR_LN2
r. ADDR_CITY_NM
s. ADDR_STATE_ABBRV
t. ADDR_POSTL_CD
u. HOME_PHONE_CD v. WORK_PHONE_CD w. EMAIL_ADDRR

Requirements Details:

Narrative: Overview

In order to identify the set of patients to include. For both Meds and Diagnoses start with groupings of them that are acceptable for inclusion. Then identify the complement of the include meds and DXs. From the complement of the include meds and DXs identify subjects who should be excluded. Subtract the exclude subjects from the population of subjects who are not excluded. These are the healthy subjects.

Diagnosis

From a set of individual or range of ICD10 DXs identify Healthy/Include ICD10 DXs. Crosswalk the ICD10 codes to a set of Healthy/Include ICD9 DXs. Subtract the Healthy/Include ICD10 and ICD9 DXs to derive a set of Exclude Diagnosis. Identify Patients who have any Exclude DXs as Admit, Medical History, Encounter, or Primary diagnoses. These patients are to be excluded DXs.

Meds

Identify Include Medications starting from a set of medication therapeutic classes, and any medication that has been ordered where the order class code is OTC. Find the complement of the Include Meds to identify the Exclude Meds. Identify any subject who has had an Exclude Med as an Ordered Med, or a Current Med. These are patients to be excluded based on Meds.

Gross Include Patients

Identify the complement of the combination of the Exclude Med Pats, and the Exclude DX Pats.

Patient with Encounters

Identify Subjects who are between 21 and 89 who have had an encounter in that last five years (since 01/01/2011).

With the following Encounter Types:

'CONSULT-TRANSCRIBED','DISCHARGE SUMMARY-TRANSCRIBED','ED CONSULT-TRANSCRIBED','ED PROGRESS NOTE-TRANSCRIBED','H&P-TRANSCRIBED','HOSPITAL ACTIVITY','HOSPITAL ENCOUNTER','INPATIENT PROGRESS NOTE','INPATIENT PROGRESS NOTES-TRANSCRIBED','OFFICE VISIT','OFFICE VISIT - REHAB','OFFICE VISIT-ECX','OFFICE VISIT-TRANSCRIBED','PRENATAL','PRENATAL INITIAL','PROCEDURE','PROCEDURE - TRANSCRIBED','PROCEDURE-ECX','RESEARCH ENCOUNTER'

The Visit Provider is of the Provider Type:

'Physician', 'Osteopath', 'Nurse Practitioner', 'Physician Assistant', 'Midwife'

Use OHSUDW.HDW_Provider_Dim as source for provider info. This source is adjucated, and is a better source for provider. Here are some criteria applied to this source:

referral_source_type = 'Internal'

Email is from the ohsu domain (@ohsu), and it is not no_email@ohsu.edu

The encounter was in Service_Area for OHSU Visits.

Family Practice

Exclude subjects whom have a PCP where:
The provider credential department is not Family Medicine.
If the provider credential department is not populated exclude subjects whose PCP's primary department does not start with 'FM' (Family Medicine)
Additional steps were incorporated to filter out any subjects whose current PCP is on the Provider List shared by the Family Medicine Department.

MyChart Research Opt-Out
Exclude subjects whose most recent status of an hmm_aud_mod_c of 458 in the table clarity.hm_mod_aud indicates that they wish to opt-out of MyChart Research Recruitment.

Healthy Subjects
Find the intersection of the Gross Include Patients with Patients with Encounters. These subjects will have a vital status flag indicated they are Alive, and there most recent MyChart History status will be Activated.

Appendix 2

Complete list of ICD-10 ranges used for Burdick/Samuels/Beninato algorithm

ICD-10 Code and Ranges	SHORT DESCRIPTION
E65, E66	Localized adiposity, Overweight and obesity
D10 – D36, O50, E73, E86, E61	Benign neoplasm of organs and body parts
H00 – H28	Hordeolum externum and Hordeolum internum of various parts of eyelids
H43, H44	Disorders of vitreous body and globe
H49 – H52	Disorders of ocular muscles, binocular movement, accommodation and refraction
H60 – H94	Diseases of external ear, Diseases of middle ear and mastoid, Diseases of inner ear, Other disorders of ear
J00 – J39	Acute upper respiratory infections, Influenza and pneumonia, Other acute lower respiratory infections, Other diseases of upper respiratory tract
K00 – K14	Diseases of oral cavity and salivary glands
K35 – K46	Diseases of appendix, hernia
K64	Hemorrhoids and perianal venous thrombosis
K80, K81	Cholelithiasis, Cholecystitis
L00 – L08, L21 – L30, L50, L60 – L75	Infections of the skin and subcutaneous tissue, Bullous disorders, Dermatitis and eczema
M15 – M27, M65-M67, M70 – M77	Osteoarthritis, Other joint disorders, Dentofacial anomalies [including malocclusion] and other disorders of jaw, Disorders of synovium and tendon, Other soft tissue disorders.
N30, N34	Cystitis, Urethritis and urethral syndrome
N40, N43, N44, N45, N47, N48, N49, N53	Benign prostatic hyperplasia, Hydrocele and spermatocele, Noninflammatory disorders of testis, Orchitis and epididymitis, Disorders of prepuce, Other disorders of penis, Inflammatory disorders of male genital organs, not elsewhere classified, Other male sexual dysfunction
N60, N61, N63	Benign mammary dysplasia, Inflammatory disorders of breast, Unspecified lump in breast
N95	Menopausal and other perimenopausal disorders
O00 – O99	Pregnancy with abortive outcome, Supervision of high risk pregnancy, Edema, proteinuria and hypertensive disorders in pregnancy, childbirth and the puerperium, Other maternal disorders predominantly related to pregnancy, Maternal care related to the fetus and amniotic cavity and possible delivery problems, Complications of labor and delivery, Encounter for delivery, Complications predominantly related to the puerperium, Other obstetric conditions, not elsewhere classified.

ICD-10 Code and Ranges	SHORT DESCRIPTION
S00 – S99	Injuries to the head, Injuries to the neck, Injuries to the thorax, Injuries to the abdomen, lower back, lumbar spine, pelvis and external genitals, Injuries to the shoulder and upper arm, Injuries to the elbow and forearm. Injuries to the wrist, hand and fingers. Injuries to the hip and thigh, Injuries to the knee and lower leg, Injuries to the ankle and foot
T07 – T79	Injuries involving multiple body regions, Injury of unspecified body region, Effects of foreign body entering through natural orifice, Burns and corrosions of external body surface, specified by site, Burns and corrosions confined to eye and internal organs, Burns and corrosions of multiple and unspecified body regions, Frostbite, Poisoning by, adverse effect of and under dosing of drugs, medicaments and biological substances, Toxic effects of substances chiefly nonmedicinal as to source, Other and unspecified effects of external causes, Certain early complications of trauma
V00 – V99	Pedestrian injured in transport accident, Pedal cycle rider injured in transport accident, Motorcycle rider injured in transport accident, Occupant of three-wheeled motor vehicle injured in transport accident, Car occupant injured in transport accident, Occupant of pick-up truck or van injured in transport accident, Occupant of heavy transport vehicle injured in transport accident, Bus occupant injured in transport accident, Other land transport accidents, Water transport accidents, Air and space transport accidents, Other and unspecified transport accidents
Z00 – Z13, Z18, Z30 – Z39, Z55 – Z65	Persons encountering health services for examinations, Retained foreign body fragments, Persons encountering health services in circumstances related to reproduction, Persons with potential health hazards related to socioeconomic and psychosocial circumstances
Z66, Z67, Z68	Do not resuscitate status, Blood type, Body mass index (BMI)

Appendix 3

Results of RW search for healthy patients: 39 retrieved; 7 had tobacco use.
32 patients without tobacco use listed here:

Chart review	Age	Allergies	Care Plan	MyChart
Wound infection, neck cyst	21-year old	No Known Allergies	No	declined
Healthy patient	21-year old	No Known Allergies	No	activated
Rib fractures after accident but did well; healthy patient	21-year old	No Known Allergies	No	activated
Foot fracture, ankle fracture	23-year old	Sulfa (Sulfonamide Antibiotics)	No	pending
Healthy	24-year old	No Known Allergies	No	activated
Healthy no problems listed	25-year old	No Known Allergies	No	pending
Healthy	25-year old	No Known Allergies	No	pending
Healthy (chart review not possible)	25-year old	No Known Allergies	No	activated
Supervision for normal first pregnancy GBS bacterirum—healthy patient	26-year old	No Known Allergies	No	pending
Surgery on ear tubes as child--healthy	26-year old	No Known Allergies	No	activated
UTI, normal, healthy	26-year old	Doxycycline	No	pending
Dental infection, healthy	27-year old	No Known Allergies	No	activated
Wisdom tooth extraction, healthy	27-year old	Penicillin	No	activated
Knee arthroscopy, healthy, adenoidectomy	28-year old	No Known Allergies	No	declined
Healthy, meds for dysuria	28-year old	No Known Allergies	No	pending
History of chicken pox, healthy, cholecystectomy	33-year old	Penicillin	No	pending

Chart review	Age	Allergies	Care Plan	MyChart
Low back pain, healthy, vitamin D deficiency	33-year old	No Known Allergies	No	activated
Cesarean section, healthy	33-year old	No Known Allergies	No	pending
Bone cyst, jaw surgery, healthy	33-year old			pending
Assume healthy, chart not available	36-year old	Tetanus Vaccines And Toxoid	No	activated
Tear of PCL, healthy	36-year old	Azithromycin	No	pending
Exposure to TB, healthy	37-year old	No Known Allergies	No	pending
Healthy	38-year old	No Known Allergies	No	pending
Healthy	41-year old	No Known Allergies	No	declined
Assume healthy as chart restricted	41-year old	No Known Allergies	No	activated
Healthy	41-year old	No Known Allergies	No	declined
Infertility, healthy	44-year old	Tylenol (Benzocaine]	No	activated
Medullary thyroid carcinoma, long problem list	45-year old	No Known Allergies	No	pending
Partial hysterectomy, healthy, goiter	47-year old	No Known Allergies	No	pending
Tonsillectomy, healthy	49-year old	No Known Allergies	No	activated
Assume healthy, restricted	55-year old	Ibuprofen	No	pending
Ganglion cyst excision, healthy	59-year old	No Known Allergies	No	pending