University of Nebraska Medical Center
BREAKTHROUGHS FOR LIFE

University of Nebraska Medical Center
DigitalCommons@UNMC

Theses & Dissertations

Graduate Studies

Fall 8-17-2018

An Evaluation of the Use of a Clinical Research Data Warehouse and I2b2 Infrastructure to Facilitate Replication of Research

Bret Gardner
University of Nebraska Medical Center

Follow this and additional works at: https://digitalcommons.unmc.edu/etd

 Part of the Bioinformatics Commons, and the Other Medical Sciences Commons

Recommended Citation

Gardner, Bret, "An Evaluation of the Use of a Clinical Research Data Warehouse and I2b2 Infrastructure to Facilitate Replication of Research" (2018). *Theses & Dissertations*. 302.
https://digitalcommons.unmc.edu/etd/302

This Dissertation 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.

AN EVALUATION OF THE USE OF A CLINIC AL RESEARCH DATA WAREHOUSE AND I2B2 INFRASTRUCTURE TO FACILITATE REPLICATION OF RESEARCH

by

Bret J. Gardner

A DISSERTATION

Presented to the Faculty of

the University of Nebraska Graduate College

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy

Biomedical Informatics

Graduate Program

Under the Supervision of Professor James C. McClay

University of Nebraska Medical Center

Omaha, Nebraska

August 2018

Supervisory Committee:

W. Scott Campbell, PhD

Dario Ghersi, MD, PhD

Julie L. Fedderson, MD

John R. Windle, MD

ACKNOWLEDGEMENTS

My quest for knowledge during graduate school, exemplified by the work described in this dissertation, has not been a solitary effort. I have many individuals to thank for their support of my goals.

First, I am grateful for the MD/PhD Scholars Program and its directors at UNMC. Dr. Smith and Dr. Romberger patiently helped me identify my interest in biomedical informatics and facilitated an opportunity to work with Dr. McClay. Their willingness to meet and to help me find options to achieve long term goals enabled the research path I have chosen. Beyond the directors, countless peers in the program provided emotional as well as intellectual support throughout graduate school and especially on the projects described in this dissertation.

I am also indebted to Dr. McClay. He has been an excellent mentor and advisor. Despite my procrastination he finds time to review my work and offer meaningful insights. He also assisted me in starting to become an independent researcher, encouraging me to find my own interests and seek my own answers. I appreciate his support of my ideas and his patience as I navigated through this experience.

I am grateful for my supervisory committee. For the past several years they have provided meaningful feedback on my writing and insightful recommendations for advancing this research. They have supported my goals of becoming a clinician scientist, both in the work described herein and in my clinical pursuits.

My ability to achieve would have been very much limited without support of the technical team at UNMC. Dr. Campbell acted as an excellent mentor. He supported my attendance at conference and was instrumental in the design and work especially of

creating a computable phenotype for pregnancy. Jay Pedersen has been a friend and a tutor. He oriented me to SQL and to relational databases. He has selflessly given of his personal time to assist in my various projects. He was instrumental specifically in taking geocoding from a concept to a creation.

Finally, I cannot thank my family enough for their support. From an early age my parents encouraged curiosity and problem solving. They have always supported my goals, by word and action. My wife is my equal partner and deserves great credit for any of my achievements. I appreciate her willingness to edit my writing, review my ideas and plans, and support our family as I often worked abnormal hours. I also am grateful for my children whose zeal for life is contagious.

ABSTRACT:

AN EVALUATION OF THE USE OF A CLINICAL RESEARCH DATA WAREHOUSE AND I2B2 INFRASTRUCTURE TO FACILITATE REPLICATION OF RESEARCH

Bret J. Gardner, PhD

University of Nebraska, 2018

Supervisor: James C. McClay, MD, MS

Replication of clinical research is requisite for forming effective clinical decisions and guidelines. While rerunning a clinical trial may be unethical and prohibitively expensive, the adoption of EHRs and the infrastructure for distributed research networks provide access to clinical data for observational and retrospective studies. Herein I deomonstrate a means of using these tools to validate existing results and extend the findings to novel populations. I describe the process of evaluating published risk models as well as local data and infrastructure to assess the replicability of the study. I use an example of a risk model unable to be replicated as well as a study of in-hospital mortality risk I replicated using UNMC's clinical research data warehouse.

In these examples and other studies we have participated in, some elements are commonly missing or under-developed. One such missing element is a consistent and computable phenotype for pregnancy status based on data recorded in the EHR. I survey local clinical data and identify a number of variables correlated with pregnancy as well as demonstrate the data required to identify the temporal bounds of a pregnancy episode. Next, another common obstacle to replicating risk models is the necessity of linking to alternative data sources while maintaining data in a de-identified database. I demonstrate a pipeline for linking clinical data to socioeconomic variables and indices

obtained from the American Community Survey (ACS). While these data are location-based, I provide a method for storing them in a HIPAA compliant fashion so as not to identify a patient's location.

While full and efficient replication of all clinical studies is still a future goal, the demonstration of replication as well as beginning the development of a computable phenotype for pregnancy and the incorporation of location based data in a de-identified data warehouse demonstrate how the EHR data and a research infrastructure may be used to facilitate this effort.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CDM	Common Data Model
CDRN	Clinical Data Research Network
CDS	Clinical Decision Support
CER	Comparative Effectiveness Research
CONSORT	Consolidated Standards of Reporting Trials
CPT-4	Current Procedural Terminology, 4th Edition
CRDW	Clinical Research Data Warehouse
DDI	Data Discovery Index
DRN	Distributed Research Network
EDD	Estimated Delivery Date
EHR	Electronic Health Record
FDAAA	Food and Drug Administration Amendments Act
GPC	Greater Plains Collaborative
HIPAA	Health Insurance Portability and Accountability Act
HITECH	Health Information Technology for Economic and Clinical Health
i2b2	Informatics for Integrating Biology and the Bedside
ICD	International Classification of Disease
ICMJE	International Committee of Medical Journal Editors
IRB	Institutional Review Board

LMP	Last Menstrual Period
LOINC	Logical Observation Identifiers Names and Codes
NAACCR	North American Association of Central Cancer Registries
NCBC	National Center for Biomedical Computing
NeHII	Nebraska Health Information Initiative
NIH	National Institutes of Health
NLP	Natural Language Processing
NSF	National Science Foundation
OMB	Office of Management and Budget
ONC	Office of the National Coordinator
PCORI	Patient-Centered Outcomes Research Institute
PCORnet	National Patient-Centered Clinical Research Network
PHI	Protected Health Information
PhRMA	Pharmaceutical Research and Manufacturers of America
RCT	Randomized Controlled Trial
REDCap	Research Electronic Data Capture
SNOMED-CT	Systematized Nomenclature of Medicine – Clinical Terms
SSDI	Social Security Death Index
UNMC	University of Nebraska Medical Center

INTRODUCTION

Overview

Consistently improving clinical care relies on validated evidence applicable to the current patient. While any number of experimental designs or anecdotal evidence may purport to resolve the mystery, scientifically, the randomized controlled clinical trial is considered the gold standard for answering clinical questions [1]. Despite the need for high-quality evidence to shape clinical practice, Francis Collins (the director of the National Institutes of Health (NIH)) et al. note only 11% of practice recommendations issued by cardiologist specialty societies "were based on 'level A' evidence, that is, evidence based on multiple well-done randomized trials" [2]. Due to challenges of conducting clinical trials, insufficient data may be published. In addition, published data is not always well-validated. Finally, the results of replicated studies may not be extendable to a novel population. For these reasons, demonstrating an efficient method for reliably replicating and extending research remains paramount.

While these challenges exist, recent advances in technology and policy may facilitate the accrual and application of actionable data in the clinical setting. First, increased adoption of electronic health records (EHRs) provides an alternative source of clinical data for research. Next, the growing number and established infrastructure of distributed research networks (DRNs) have the potential to enable pragmatic trials and observational studies for comparative effectiveness research (CER) at a greater rate than was previously possible. Finally, these tools may be harnessed to allow not only novel research, but, replication and extension of previously published results. While exhibiting great potential, each of these tools also has inherent weaknesses. One

significant obstacle is heterogeneity between datasets being combined within distributed networks [3]. A second obstacle is consistent identification of a study population across institutions. These concerns must be accounted for and addressed in any effort to produce clinical data.

In this introduction, I review various factors contributing to the insufficiency and lack of applicability of reliable clinical data. I also explore the benefits and challenges of the use of EHRs and DRNs for clinical research. Finally, I address how these tools may be harnessed to facilitate replication research to validate previously published results and enhance the applicability of clinical data.

Data Challenges Impacting Improving Clinical Decisions

Challenges to Conducting Clinical Trials

As noted above, there is a dearth of high-quality and applicable evidence for creating and updating clinical guidelines. This may in part be due to the barriers to conducting clinical research. These challenges exist at all elements of the research process, from planning a study to implementing changes based on results. Challenges may be grouped into categories of trial administration and cost, recruitment, and publication. All of these categories must be fully addressed for a trial to succeed.

Resource and Time Requirements for Trial Administration

Trial administration requires investment of a great deal of resources in terms of time, personnel, and money. Eligibility screening usually occurs before consent and entails reviewing medical records, often necessitating obtaining additional information from other institutions and providers [4]. Penberthy et al. reviewed cancer trials over an 18-month period and discovered that three to thirteen patients were screened for each

patient actually enrolled at a cost of $129 to $336 per enrolled patient [4]. Additionally, they noted that more than 50% of the screening took 30 or more minutes per patient [4]. While resource intensive, eligibility screening represents only one small element of a clinical trial.

Many additional costs exist in conducting clinical trials. In 2010, of the $46.4 billion spent by Pharmaceutical Research and Manufacturers of America (PhRMA) member companies on R&D, $32.5 billion went toward clinical trials [5]. Xu et al. estimate that it costs over $800 million to develop a new drug [6]. Costs for reimbursing patients, paying for time to navigate the complex web of the Institutional Review Board (IRB) approval process, and translating questionnaires or other documents are just a few of the many expenses likely to be incurred when conducting a clinical trial. Emanuel et al. concluded that on average, 200 hours were required per patient for pharmaceutical industry-sponsored trials. 32% of these hours were for non-clinical activities such as those noted above. They concluded that, “on average, excluding overhead expenses, it cost slightly more than $6,094 [. . .] per enrolled subject for an industry-sponsored trial, including $1,999 devoted to nonclinical costs” [7]. Berndt and Cockburn explain that, “the growing complexity of clinical trials and of the underlying science suggests that more time, more highly trained personnel, and more sophisticated equipment may be required to conduct a typical study [5]. The U.S. Bureau of Labor Statistics estimated that the input costs for conducting clinical trials rose 8% per year between 1989 and 2011, nearly double the inflation rate in the NIH Biomedical R&D Price Index during the same period [5]. With today’s extremely competitive funding environment, such costs may be insurmountable for a single researcher or research team.

Recruitment

The next set of challenges in conducting effective trials deals with recruitment. The population recruited must be large enough for statistical analysis, representative of the population results will be applied to, and reachable by the researchers conducting the trial. Califf et al. note that 62% of interventional trials registered on ClinicalTrials.gov between 2007 and 2010 had fewer than 100 participants [8]. Additionally, they described that the majority (66%) of registered trials were single-center [8]. Bernardez-Pereira et al. studied over 7,000 cardiovascular clinical trials from 2000 through 2013 and concluded low recruitment represented the primary cause of early termination [9].

Connected to low recruitment numbers is the difficulty of generalizing the results of clinical trials. Researchers estimate that only 2% to 3% of adult cancer patients are enrolled in clinical trials in the United States [10]. Furthermore, minorities are often underrepresented relative to the portion of cancer patients composed of any given minority [11]. For instance, Blacks have a higher incidence and mortality rate for most cancers compared to Whites, nearly double for Black males relative to White males for prostate cancer [11]. However, from 1998 to 2001 the total number of Black patients enrolled in National Cancer Institute sponsored clinical trials increased by only 38 while the total number of participants increased by more than 6,500 [11]. Such underrepresentation makes it difficult if not impossible to extend results of carefully constructed clinical trials to all cancer patients. Additionally, potential population specific polymorphisms effecting pharmacokinetics or pharmacodynamics may not be recognized [11]. With such a small percentage of patients being enrolled in trials, and this small cohort not always being an accurate representation of the entire patient population, it is difficult for generalize results from these studies.

With the cost and resource requirements to conduct clinical trials compounded by the challenge to recruit a patient population representative of the target population,

alternate strategies must be employed to effectively acquire and validate high-quality clinical data.

Extant Data Not Validated

Overview

Despite these challenges, researchers persist and clinical data continues to be published. ClinicalTrials.gov reports in 2000, fewer than 4,000 trials were registered. By 2015 this number has exceeded 180,000 (http://ClinicalTrials.gov). This rise in the number of registered trials likely represents both a total increase in the number of trials being conducted as well as an increased proportion of trials registering. Policy changes may influence the proportion registering. In September 2005 the International Committee of Medical Journal Editors (ICMJE) required registration as a condition of publication. In December 2007 the Food and Drug Administration Amendments Act of 2007 (FDAAA) was passed by congress, expanding the types of trials to be registered, increasing trial registration information to include summary results and adverse events, and imposed potential penalties for non-compliance including civil monetary penalties or withholding of NIH grant funding [12].

Despite this increase in clinical trials, data concerns remain. Recent emphasis has been given to encouraging improved data management for accessibility and reuse [13]. However, many studies go unpublished or are published after a great delay (median of 21 months between completion of trial and publication in a journal) [14]. Few studies are replicated or independently validated. This is due in part to publication bias, increased emphasis on patient privacy, the aforementioned challenges to conducting clinical research, and a culture favoring rapid output of novel results.

While clinical data has been accumulating for millennia, with the challenges to conducting original research, much of these data have not been validated [15, 16]. This trend continues for studies relying on data extracted from the EHR [17]. In a systematic review of risk prediction models based on EHR data, only 26 of 107 studies performed validation across sites [18]. In addition to some studies without evidence of attempting to be validated, examples abound of observational studies contradicting previous studies or of being rejected by randomized trials performed later [19]. Some instances of such contradictory results include studies of the relation between hormone replacement therapy (HRT) and cardiovascular risk [20, 21], bisphosphanate use and cancer risk [22, 23], and fracture risk accompanying use of statins [24-26]. Wagenmakers et al put the situation in perspective stating “findings that do not replicate are worse than fairy tales” [27]. In light of these many contradictory findings, accepting any single study becomes questionable.

With the potential for contradictory results, working toward replication is paramount. In this section, I explain the importance of reproducibility in science, review disparate definitions of reproducible research, explore the impact of the paucity of reproducible results, examine obstacles to replication studies stemming from publication practices as well as local infrastructure challenges, and describe some of the existing guidelines in various fields for moving toward reproducible research. While efforts are underway in many fields to enable and encourage reproducibility, this paradigm shift will take time.

Importance of Reproducibility in Science

Reproducibility has been referred to as the cornerstone and Supreme Court of science, and as the best and possibly the only believable evidence for reliability of an effect [28]. Wagenmakers et al state “findings that do not replicate are worse than fairy tales” [27]. Cacioppo continues, “Reproducibility is a minimum necessary condition for a

finding to be believable and informative" [29]. Despite the essential nature of replication, across all domains of science, many studies are not validated [30, 31]. Reliably replicating results with consideration of resource limitations is essential for evidence based medicine.

Mixed Results from Historic Replication Efforts

Lack of validation may be the result of there being no attempts made to replicate a study or the failure of attempts to replicate published results. In some cases, replication is not attempted or published. As noted earlier, a systematic review of risk prediction models based on EHR data revealed only 26 of 107 studies performed validation across sites [18]. Even when replication is attempted, lack of reproducibility has not been seen to correlate with journal impact factor, the number of publications about a specific finding, or the number of collaborators on the publication [32]. In one review, only one-third to one-half of studies in high-ranking psychology journals demonstrated replicable results [16]. In the pharmaceutical industry, target validation is often attempted before advancing a potential drug to phase II trials. In an industry sponsored study, despite basing this validation on published material, key data were only reproduced in 25% of cases [32]. This low rate held true even when cell lines or assay formats were modified. The results which were reproducible were demonstrable irrespective of modifying the experiment environment, while the majority of results were not reproducible whatever the experimental conditions. A similar set of experiments revealed only 11% of pre-clinical oncology results could be replicated [33]. These contradictory results may stem from errors or multiplicity in the original study, errors in the attempted replication, insufficient information in the publication to facilitate replication, insufficient infrastructure or data at an independent site to replicate the study, or a study demonstrating an effect that may exhibit reproducibility but not replicability. The inability

to replicate a study calls into question the published results and makes it challenging to have confidence in clinical guidelines stemming from this work.

Definitions

In order to promote and adopt practices toward reproducible research, a clear definition of reproducibility is essential. Despite being a priority in computer science [34-37], bioinformatics [38], biostatistics [39], epidemiology [40], and clinical trials [41], as well as being promoted by the NIH [42], no common, agreed upon and accepted definition of reproducibility is currently in use across scientific disciplines [30, 43].

Historically, Claerbou defined reproducible research within computer science as providing sufficient code and original data to allow the reader of the publication to be able to view the entire process from raw data collection to publication of results [44]. Other researchers, however, define reproducibility as "the provision of sufficient methodological detail about a study so it could, in theory or in actuality, be exactly repeated by investigators" [43]. Further confusing matters, the terms reproducibility, replication, duplication, and generalization are often used interchangeably or inconsistently in publications.

For the remainder of this work, I will adhere to definitions recommended to the National Science Foundation (NSF) [29] (Table 1). Herein, reproducibility "refers to the ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator" [29]. While replicability is "the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected" [29]. Goodman et al help clarify the distinctions between reproducibility and replicability. They synonymize reproducibility with "methods reproducibility" [30]. In essence, this consists of the original researcher providing raw data and all source code and software to a novel investigator. The novel investigator may

then use the same data and same programs to attempt to generate the same results. Contrarily, replicability is explained as “results reproducibility”. In this case, the original researcher provides a detailed protocol and clear explanation of results to a novel investigator. The novel investigator may then conduct as similar an experiment as possible with new data collection in a new environment to see if the same results are produced. Thus reproducibility serves to validate the methods and replication offers a second witness of the observed effect.

Term	Synonym	Definition
Reproducibility	methods reproducibility*	refers to the ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator**
Replicability	results reproducibility*	the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected**

Table 1. Definitions of Reproducibility and Replication. *[30] , **[29]

Obstacles to Replication – Publication and Local Infrastructure

Many challenges exist hindering researchers' ability to replicate studies. These challenges occur both in the current publication environment as well as in the local infrastructure researchers work in. The current publication environment suffers from publication bias, has no systematic means of publishing sufficient data and protocols, and has no consistent metric to assess the replicability or reproducibility of a study. The peer review process is not able to detect all errors [45], and rejected manuscripts are often published in another location [46, 47]. While irreplicable results are rarely the result of fraudulence [48], obstacles continue to impede the publication of replicable work.

Publication Bias:

Many studies over the past decades demonstrate a bias in published journals toward studies showing positive results [14, 49-51]. This "publication bias" as well as delay in publication (median of 21 months between completion of trial and publication in a journal) limits clinician access to complete data, may lead to duplicated research, publication of spurious results, and limits improvements to the quality of medical care implemented [14, 52, 53]. In addition, this publication practice may lead to the "file-drawer" problem where significant, negative results are tucked away privately and Type I errors are published [54]. Future clinical guidelines and clinical trials will thus be based on incorrect conclusions. Publication bias may also promote multiplicity [30]. Multiplicity is the practice of testing many hypotheses, multiple associations, innumerable models, endpoints not defined *a priori*, hypothesizing after the results are known (HARKing), and often failing to publish this winding protocol which eventually led to a significant result [55-63]. Many negative results lost in file drawers may replicate well if they were published. Type I errors have a low likelihood of replicating no matter how carefully a protocol is observed. Effects called significant after irregular statistical analyses will likely fail to be

dubbed significant when a single, standard analysis is performed on novel data [64]. These issues are exacerbated by some of the aforementioned challenges in clinical research. For instance, a small sample size leads to low statistical power and may hamper a researcher's ability to demonstrate statistical significance with traditional analyses.

Data and Protocol Sharing:

In addition to publication bias leading to the dissemination of non-replicable studies, the current publication environment has no systematic means of sharing detailed protocols and raw data for EHR based studies [65]. This is due both to current systems as well as culture. Currently, biomedical journals have no unified approach for authors to submit code or software used to extract, transform, or analyze data. Additionally, no uniform repository or approach exists for making raw data available. Culturally and legally, if such a repository were made available, patient privacy for EHR based studies may preclude sharing of data from any point in the pipeline. Researchers desiring to replicate or reproduce prior studies are haltered by this lack of consistent sharing of code or data.

Next, researchers are further limited due to inconsistent publication of written descriptions of protocols. While clinical research may involve extremely detailed phenotypes and inclusion and exclusion criteria to define the cohort of interest, studies have demonstrated the practice of incomplete protocol or phenotype publication. In a review of research studies using UK EHR data, only 5.1% published the entire set of terms sufficient to implement the EHR-derived phenotype [66]. Even if listed, EHR-based phenotypes are usually published as human-readable, complex documents rather than machine-readable programs [67]. Also, these complex documents of innumerable terms often still lack sufficient context, such as noting if the diagnosis being queried was the

primary cause of admission [43]. In a separate review, only one of 400 biomedical studies published a complete protocol [68]. With the diversity of proprietary formats for EHR data collection and storage, research pipelines involving data extraction, pre-processing, and manipulation are similarly diverse. However, associated programmatic code is rarely published [43]. The combination of a lack of a systematic approach to sharing programmatic code or data coupled with a culture of publishing incomplete protocols leads to published studies that will not be replicable. To facilitate reproducibility, many studies developing risk models report model coefficients [18]. There is no systematic approach to this practice and the calculation of such coefficients may remain a black box. Independent researchers are left to interpret the limited methods and may not be able to conduct similar analyses or identify a similar population.

Cultural Limitations:

Finally, in addition to publication bias and lack of infrastructure or culture for publishing sufficient methods and data for reproducible research, there is no systematic metric in place to assess the replicability of a published study or reward authors for adhering to replication practices. Journals could incentivize publication of raw data and programmatic code, note if authors pre-registered hypotheses, indicate if all co-authors have reviewed and have access to the raw data, or otherwise note via reviewer feedback if the publication provides sufficient information for replicability [69].

Policy Changes Increasing Replicability of Published Studies

Recently, approaches have been proposed for overcoming the aforementioned obstacles to replication. First, clinical research has had safeguards in place for some time to protect patients and maintain a high level of scientific integrity. These safeguards are not as developed or mandated in pre-clinical or basic science research. When human subjects are involved, a detailed protocol, including justifications for sample size

and planned analysis, must be submitted to and reviewed and approved by a local IRB. Being required to state hypotheses and analysis approach before any patient data collection will prevent at least some multiplicity. Modifications are possible and wandering statistical analysis may still occur prior to final publication of the results, however, the checkpoints currently in place promote replicable studies. Reporting of some of these details is outlined in the revised CONSORT (Consolidated Standards of Reporting Trials) guidelines [70]. Herein researchers should ensure they report on the 22 points of a standardized checklist as well as clearly outline the experimental design in a flowsheet detailing enrollment, intervention allocation, follow-up and analysis. In this way, other researchers may evaluate the scientific merit and reliability of the results being published.

Next, the FAIR (Findable, Accessible, Interoperable, Reusable) Data Principles [13] set forth by a diverse set of stakeholders from academia, industry, publishers and funding agencies, outline the desiderata of reproducible and replicable science. While not prescribing specific structure or location for data sharing, these principles provide clear definitions and represent the cultural recognition of the need for revisions. The authors clearly relate the interoperability and openness of data to scientific advancement and make plain the necessity of having data and protocols both human and machine readable. Furthermore, example repositories for diverse data types are noted as starting points for seamless data sharing. As these principles are adopted and as a greater infrastructure is developed for standard data sharing in each scientific domain, replication will be enabled to a much greater degree.

Next, in 2014, Francis Collins announced plans the NIH was implementing toward reproducibility and replication in science [42]. These plans were designed to provide better training toward replicable science as well as provide forums for scientific

discussion and repositories for raw data. For researchers, a training module was to be developed emphasizing reproducibility and good experimental design. For reviewers of grant applications, a systematic checklist was proposed. This checklist would include an evaluation of the planned analyses, a review of the antecedent work the current grant is based on, and may necessitate replicating previous studies if their results were questionable. The NIH plan further called for a Data Discovery Index (DDI) where researchers could locate and access unpublished, primary data. With this repository, the new researcher could then cite the data source if used thereby giving credit for original data collection. Finally, Collins proposed the development of PubMed Commons, a forum for open discourse about scientific articles. Collaborations, clarifications, questions, and criticisms for published work could be documented at this site. These plans may help shift the paradigm of researchers and other stakeholders, elevating the importance of publishing replicable results and rewarding the replication of findings.

Additional guidelines are being set forth from a variety of disciplines. In 2007 the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) [71] guidelines outlined what should be included for an accurate and complete report of observational studies. As clinically captured data have become more available and abundant, further recommendations were made in the REporting of studies Conducted using Observational Routinely collected health Data (RECORD) statement [72]. This 13 point checklist outlines recommendations for reporting on the type of study, provenance of data, data cleaning and analysis, any data linkages, and the location and time of the study as well as a description of the participants, results, generalizability, and accessibility of protocol and code. These guidelines are available to researchers, editors, and other stakeholders. They provide examples of good reporting, though, no consistent standard is prescribed.

With these cultural and policy advances toward increased replicability of published results, an effective means of conducting clinical research and replicating published results must be developed. With the challenges described above, it is not feasible, and may not be ethical, to replicate clinical trials at novel sites. However, validating and extending existing results is critical to improve patient care. The increased use of electronic health records in conjunction with the growing number and infrastructure of distributed research networks may provide a framework for collecting clinical data and independently replicating clinical results.

Potential Solutions for Making More Data Actionable

Increased Adoption of Electronic Health Records

Historically, paper charts were the repository for patient information. These bulky files included quickly scrawled clinical impression and were stored at the site of care. In recent years, electronic health records have replaced this archaic system. For instance, in 2001, only 18% of office-based physicians utilized any type of EHR [73]. Stimulated by President Bush's 2004 Executive Order 13335, which created the position of National Coordinator for Health Information Technology, and incentivized through the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009, in 2013 at least 78% of office-based physicians had implemented an EHR [73, 74]. Private acute care hospitals saw similar increases in EHR adoption, rising from 9% in 2008 to 59% in 2013 and 75% in 2014 [75, 76].

Clinical Benefits of Electronic Health Records

Electronic capture and storage of patient data is not merely a new approach to record keeping, rather, EHRs offer benefits toward interoperability, patient safety, and

efficiency. Interoperability across the healthcare community is becoming a national goal. In 2015 Karen B. DeSalvo, MD, MPH, MSc, the National Coordinator (ONC) for Health Information Technology, published “A Shared Nationwide Interoperability Roadmap” outlining critical actions both the public and private sector of healthcare must adopt to enable an interoperable ecosystem of health IT within a decade [77]. As vendors and the healthcare community adhere to the ONC standards in the creation and use of EHRs, the ability to share clinical information across institutions and care settings will come to fruition.

As interoperability is achieved, advances in patient safety are enabled. With standards employed, clinical decision support (CDS) systems may be centrally developed and tested and then applied across diverse EHRs. These CDS systems evaluate drug-drug interactions, may alert the health care team to abnormal or extreme lab values or vital signs, and may recommend the most efficacious therapy for the current condition. Historically these were developed ad hoc or required onerous adaptation at each site. Adhering to interoperability standards further advances patient safety as clinicians may more readily obtain a complete patient history. Rather than bulky charts being lost or requiring excessive time to sift through faxed copies, EHRs allow electronic transfer of this critical, clinical information. In the past, the “curly braces” problem has prohibited efficient sharing of data across sites [78]. Some of this concern is resolved as sites follow ONC guidelines. Clinicians may then have access to a more complete patient history to allow them to make the best therapeutic choice. In the future, patient safety may also be enhanced as interoperable EHRs are used for clinical research. Access to patient information across sites will enable pragmatic trials and retrospective studies to be conducted leading to greater clinical evidence. In each of these ways, interoperable EHRs move health care toward a safer patient experience.

Finally, adoption of EHRs enhances efficiency. While a handwritten chart exists in only a single physical location, EHRs allow patient information to be accessed and updated irrespective of geographic location. A single copy of a paper chart may be reviewed by only a single individual at any one time. EHRs allow all members of the health care team to review information simultaneously. Reviewing paper charts may necessitate examining many pages of scrawled notes. EHRs allow for electronic searching, data summary, calculations, and other heuristics to be employed quickly. Efficiency is increased with the use of EHRs as geographic limitations are removed, multiple individuals are able to collaborate simultaneously, and technological advances for data interpretation and searching are employed.

Clinical Research Benefits of Electronic Health Records

Overview

While the re-use of EHR data is rife with challenges, use of EHRs for comparative effectiveness research may decrease the cost of time and resources spent on eligibility screening and recruitment, allow study of a more representative population, incorporate more diverse data elements, and increase efficacy of trials that are underway. Payne et al. conducted a survey of IT use in clinical translational research and noted specific benefits of utilizing EHRs [79]. These researchers note that the data contained in EHRs may be used at many steps of the clinical trial process, including hypothesis formation, recruitment, analysis, and especially data collection [79]. Similar to records for clinical care, historically, data collection for clinical trials has depended on paper forms, with data extrapolated and codified manually [79]. These paper forms must then be transcribed into databases for analysis, a process introducing further opportunity for errors [79]. The increased use of time and resources as well as the introduction of

greater potential for human errors are paramount. Payne et al. report that harnessing EHR technology has significantly decreased time-consuming and redundant data entry and simultaneously demonstrated increases in data quality [79].

Eligibility Screening

Furthermore, recent work has shown with high fidelity electronic records, eligibility screening may be largely automated. Ni et al. demonstrated this principle and noted a 92% reduction in workload and a 450% increase in trial screening efficiency [80]. While the use of EHRs aids in screening, efforts to establish automated screening must be made at each participating site with unique EHRs.

Patient Population

In addition, data from operational EHRs compliments information gathered from well-controlled clinical trials. First, EHRs include data on an entire patient population, not a highly selected cohort. Data is collected on this population continuously rather than only at a few time points during a year. In contrast to the very limited population with strict adherence to study protocol, EHRs offer data on a very broad population that reflects actual care and patient conformance to prescribed therapy [81]. Also, within a trial only a small subset of data elements are collected and analyzed. In contrast, within an EHR data is diverse, not focused on a single disease or known risk factors. In this way, EHRs provide an abundance of data elements for the formulation of novel hypotheses and foundation for analyses. While not the same, with careful consideration, data gathered from an EHR may be valid to answer clinical research questions.

Finally, as EHRs are used for research, translating results into decision support or deploying risk calculators in clinics is much simpler than historic methods [18]. It is plain that EHRs may be used for secondary purposes of data acquisition and maintenance for

clinical research with meaningful impact on diminishing time and expense of conducting trials. However, EHR data is not collected for research – potentially leading to problems with quality, completeness and comprehensiveness of these data.

Concerns with Secondary Use of EHR Data for Clinical Research –

Whether on paper or electronically, data is recorded in the medical record by many members of the health care team, by equipment, and in some instances by patients. While the volume of data abound, there are concerns with use of these data directly for clinical research. Data captured in the EHR are recorded for clinical and billing use. As such, these data may be inaccurate, incomplete, transformed locally, lacking necessary granularity, or irreconcilable with approved research protocols [82]. Each of these potential limitations must be taken into consideration prior to drawing conclusions from analysis of these data.

Inaccurate Data:

Numerous studies evaluating data from diverse EHR vendors and locations have demonstrated spurious data recorded for patient care [82]. Extracted data may include pregnant males, treatments apparently prior to birth or following death, and patients being emergently intubated only to be released a short time later [83, 84]. Without quality analysis in place, inappropriate conclusions may be drawn from secondary use of EHR data.

Incomplete Data:

Incomplete data exists at various levels. First, a patient rarely receives all medical care throughout his/her life at a single institution. Whether traveling, moving, or changes in management of local institutions, a patient is likely to have clinical data maintained at a plethora of medical centers and clinics. Nasir et al have shown 19.1% of

readmissions for heart failure patients were to a different hospital than the original [85]. Similarly, Bourgeois et al found 31% of patients in Massachusetts visited two or more hospitals over a five-year period [86]. This fragmentation means no single clinic will be able to access a complete medical picture for the patient. In addition, as data are recorded for clinical care, only necessary studies are conducted and pertinent findings recorded. While vital signs may be recorded at nearly every visit, complete visual examination findings may not be found during a visit for a appendectomy, even if the patient is suffering from diabetic retinopathy at the time. This may lead to censoring, a property where events outside the dates of observation are absent or ambiguous [87]. The date of recorded diagnosis in the EHR is rarely the date of onset of the disease [82]. Next, EHRs are inherently limited on data from healthy individuals and biased toward those needing medical care. Finally, as hospital and clinical environments differ between institutions, the normal labs and findings regularly recorded will differ in like manner. Various elements may be routine at one institution and routinely absent at a neighboring medical center. Managing missing variables and accepting the absence of some findings is requisite for research involving EHR data.

Transformed Data:

Data in the EHR is recorded for clinical care and later processed for administrative and billing purposes. In many instances, all of the data is not available to the researcher. Claims data may be given in lieu of full access to source data. As data is processed, data quality for research may decline. First, studies have shown [88] prognostic indicators, patient reports, and disease burden are not well represented in billing data. In addition, unintentional errors may creep in through the administrative process. These may be due to coding errors, inexperience, insufficient clinician oversight, upcoding to maximize payment, underreporting quality measures, and poor

verbal or written communication [89]. Beyond these errors, standard coding has limitations and perils for research. Semantic drift may occur over time wherein certain diagnostic or other codes and vocabulary change, making longitudinal studies difficult to interpret [90]. In addition, relying solely on diagnostic codes may drastically underrepresent the study population of interest [91]. Extensive data are often retained in clinical, pathology, radiology, or other notes requiring sophisticated natural language processing (NLP) techniques to extract them with unknown efficacy [92]. Which data are available and the processing of these data within the EHR must be considered when planning and conducting any research study relying on such patient information.

Data Provenance:

Data stored in the EHR may originate from a variety of sources, including any member of the health care team, equipment or instruments, the laboratory, pathology reports, radiology reports, and the patient himself. In some instances, multiple sources may provide insight into a single issue. Administration of a medication is a prime example [82]. A patient may report what medication he is taking and what was administered during his stay. These data will be stored in the medication reconciliation form. There are orders for medications, dispensing records from the pharmacy, and the medication administration record. Each of these elements is stored in the EHR and none of these elements is definitive to the question of what medication was physically given to the patient. Selecting any one of these as a proxy for the fact of interest has implications that must be reconciled.

Data Lacking Sufficient Granularity:

While data recorded in the EHR is extensive and varied, research protocols may require data elements not present. In a review of 98 outcomes studies utilizing EHR data, 55%

of these studies supplemented EHR data with non-EHR data [93]. In addition, as noted earlier, some data made available to researchers represent only a summary of the patient's condition [82]. A plethora of clinical findings and laboratory values may contribute to the diagnosis made by a clinician. The final ICD code may be all the researcher has available. Individual elements contributing to the diagnosis are masked from the researcher and may inhibit drawing meaningful conclusions from EHR data.

Increased Developments of Distributed Research Networks

Overview

As described above, with considerations of the limitations and potential bias introduced by re-using clinically collected data, EHRs may prove a valuable tool for comparative effectiveness research. This resource may be bolstered as it is used in conjunction with distributed research networks (DRNs). While EHRs contain a diversity of data, these data are still limited to a single clinic or academic institution. Studying rare diseases or small effects may remain outside the potential for a single researcher. DRNs offer an opportunity for collaboration, data sharing, and federated queries to maximize sample size and generalizability of the results. Many of the challenges to clinical research noted earlier may be addressed through DRNs. A variety of DRNs have developed in recent history Observational Medical Outcomes Partnership (OMOP) launched in 2008 [94], Sentinel System (FDA drug monitoring system with 28 collaborators) launched in 2008 [95]. I will focus on the National-Patient Centered Clinical Research Network (PCORnet) to describe the potential benefits for comparative effectiveness research.

Example of a Distributed Research Network – PCORnet

Collins et al. describe “PCORnet aims to build a national research network, linked by a common data platform and embedded in clinical care delivery systems. This network will enable studies, and in particular randomized trials, that have been impractical to conduct to date—and do so with economies of scale” [2]. These leaders summarize the potential benefits of a massive network of linked EHRs stating, “a network of electronic medical records representing over 100 million covered lives will make large-scale observational and interventional trials faster to launch, more representative of diverse real world populations, and capable of providing much-needed answers to comparative effectiveness research questions with greater accuracy” [2]. Additionally, they note that PCORnet will facilitate conducting these effective trials at “affordable cost” [2]. With the tremendous promise PCORnet offers, as with any intervention or prescribed behavioral change in medicine, evaluating its efficacy is essential.

The Greater Plains Collaborative

In 2013, the Patient-Centered Outcomes Research Institute (PCORI) funded 11 clinical data research networks (CDRNs) and 18 patient-powered research networks (PPRNs) that compose PCORnet [96, 97]. Fleurence et al. describe the commitments for each CDRN including building a large patient cohort with comprehensive clinical data, developing necessary policies to ensure patient privacy and data security while allowing for data sharing and participation in multi-network randomized trials and observational studies, and developing policies for data standardization [97]. This system of multicenter research allows many advantages, including “greater sample size and power, the ability to study effects of practice pattern and treatment variation, the inclusion of diverse populations, and the possibility of supporting analyses that assess heterogeneity of treatment effect” [97-99]. Efficient multi-site analyses are facilitated by the creation of a

common data model (CDM) that allows for queries to be shared rather than large amounts of patient data being transferred. Developing the infrastructure for this research has been ongoing (Phase I). Recently, PCORI awarded Phase II funding to ensure sustainability and continued progress for three additional years.

The Greater Plains Collaborative (GPC) is one of these 11 CDRNs. The GPC encompasses over 10 million patients from over 20 hospitals, 700 clinical locations, and 8,000 providers for all levels of care [100]. The patients included embody a diverse mix of rural populations and urban communities, including often underrepresented minorities. The GPC institutions are committed at the highest levels to this collaborative work. As evidence of this commitment, the GPC has established a master data-sharing agreement as well as reciprocal IRB agreements to allow collaborative research queries to be run and facilitate multi-site studies. Additionally, the GPC contains a robust collaborative team of informaticists. Teams from each participating site have worked to load data into the vendor neutral informatics for integrating biology and the bedside (i2b2) data warehouse. Data sources are diverse and range across commercial EHRs including Epic and Cerner, tumor registries from the North American Association of Central Cancer Registries (NAACCR), the Social Security Administration's Data Master File for mortality, Health Information Exchange data such as NeHII and the Indiana Health Information Exchange and patient-reported outcomes via Research Electronic Data Capture (REDCap) surveys [100]. While the availability of such data has great potential for research, Waitman et al. note that application of such data to determining clinical effectiveness has not been convincingly demonstrated to date [100].

i2b2

As noted above, the GPC is utilizing Informatics for Integrating Biology and the Bedside (i2b2) to achieve its goals of collaborative clinical research. I2b2 is operational

at more than 60 academic medical centers, HMOs, and private companies. This software resides in the public domain (http://www.i2b2.org) and allows for immense individual adaptation and development while maintaining interoperability with a common messaging protocol that allows communication through web services and XML [101] . The goal of i2b2 is “to provide clinical investigators broadly with the software tools necessary to collect and manage project-related clinical research data in the genomics age as a cohesive entity, a software suite to construct and manage the modern clinical research chart” [101] . This suite allows integration of a variety of data types and allows queries to be created at one site and shared with many locations to probe locally secure data.

The University of Nebraska Medical Center (UNMC) clinical research data warehouse (CRDW) is an IRB approved environment built on i2b2. Patient-centric data is extracted from affiliated health care organization electronic health records (EHR), combined with national registries, mapped to Office of the National Coordinator (ONC) designated terminology standards and fully de-identified.

With data extraction and curation efforts at a plethora of sites, PCORnet seeks to bridge the gap to this secondary use of clinically captured data by tapping into this torrent of information from the EHR. A variety of standard and proprietary coding systems are used to organize this massive dataset. While information about medications prescribed, hospital admissions and diagnoses, family history, allergies, complications, and even genomic information may be stored, querying this data in a meaningful way for clinical research is often prohibitively cumbersome. Querying large patient sets spanning multiple institutions has historically been onerous due in part to differences in local coding and representation of facts. With the CDM and an infrastructure for querying it as well as widespread use of i2b2, federated queries for large patient sets are becoming a reality.

Persistent Challenges

While this pooling of data under a common umbrella may appear to be the panacea for obstacles encompassing clinical research, recent studies from other collaborative groups demonstrate heterogeneous data sets may yield disparate results despite application of identical methods [3, 102]. Heterogeneity of treatment effect and patient differences may be masked as observational studies amalgamate large populations [103, 104]. Significant, yet opposite, effects from distinct populations participating in the network may average out to no effect or an effect in one direction. Overhage et al. demonstrated the efficacy of using the Observational Medical Outcomes Partnership (OMOP) common data model across multiple sites and maintaining data integrity to investigate safety surveillance, however, they made no inquiry as to individual site result variations [105].

One critical, and often overlooked element of the design process, is the selection of an appropriate data source to query. Selection of a data source may often be made on the basis of convenience or accessibility rather than on any standard guidelines or rationale [19]. While attaining necessary statistical power and having data elements needed to ask the desired questions are critical elements of a well-designed observational study, recent studies demonstrate that the choice of database plays a major role and that applying the same study methods to different datasets may yield disparate results [3].

As with the use of EHRs for clinical research, adopting federated queries in a DRN has challenges that must be considered. The impact of heterogeneity of data and the effect of transforming data to a common data model require further investigation. Despite these potential issues, the infrastructure and participants in a DRN offer valuable solutions to many of the challenges I have described regarding comparative

effectiveness research. Sample sizes may be dramatically increased. Costs for large-scale trials may diminish drastically. The population and the data captured as data are populated from the EHR will be real-world rather than a potentially biased trial environment. With the ability to conduct trials, it will be critical to ensure the validity of both novel results as well as previously published material.

Leveraging EHRs and DRNs to Replicate and Extend Clinical Research

Addressing Challenges to Replication Research

Use of EHR data for observational and retrospective studies provides a potential alternative to fully duplicating clinical trials to assess the original hypothesis [106, 107]. Tannen et al demonstrated the potential to use observational studies from operational EHRs to successfully replicate the findings of a series of cardiovascular clinical trials [106]. Observational studies are increasing in popularity. During the 1990s, nearly 80,000 observational studies were published. In the following decade, this number rose to 263,557 [108]. Overhage et al note that well designed observational studies may yield effect estimates comparable to those reported in randomized controlled trials (RCTs) [19]. However, not all studies are well designed and may produce spurious or conflicting results. Young boldly states “Any claim coming from an observational study is most likely wrong” [108]. While this may be hyperbole, as with any scientific claim, the results of observational studies or clinical trials should be replicated to enhance validity.

Remaining Gaps in this Process

With the availability of clinical data extractable from widespread EHRs coupled with the infrastructure and collaborations made possible with growing distributed research networks, replication of studies may be accomplished more efficiently. The

challenges of excessive cost and time requirements, the difficulty in enrolling a sufficiently large and sufficiently diverse study population, the obstacle of data collection and storage, and the hurdles of selecting an appropriate data source and maintaining patient privacy may in large measure be met through these new tools. As these challenges are overcome, efficiently replicating a study becomes feasible.

As described earlier, the NIH and multiple scientific journals are calling for cultural and policy changes toward reproducibility and replicability in science, including clinical research. As repositories are developed to make data and protocols available and as providing these data becomes expected rather than the exception, replication of studies may increase. Despite these changes, no metric for evaluating an articles replicability nor a metric for assessing a local infrastructure have come to light.

In this dissertation, I first evaluate UNMC's CRDW infrastructure as a potential tool to rapidly replicate risk models. I explore the limitations resulting from incomplete protocol publication as well as the requirements for data volume and variety at the replication site. With the identification of missing data in terms of a computable phenotype for pregnancy and variables related to a patient's socioeconomic status, I explore the development of such a phenotype and demonstrate the integration of extra-EHR variables into the CRDW. In this way, I demonstrate and approach and considerations researchers seeking to replicate studies may use as well as provide a means for incorporating census data with clinical data in a queryable fashion. While this dissertation cannot address all outstanding obstacles and concerns relevant to conducting replication research, this work does demonstrate replication and provides additional resources to assist other researchers in this effort.

CHAPTER 1 – ATTEMPTED REPLICATION OF A READMISSION RISK MODEL FOR HEART FAILURE PATIENTS

Introduction

Clinical research results are being published at an unprecedented rate. A majority of these studies, however, involve relatively few patients and are completed at a single institution [8]. With minorities often underrepresented and such a small sample of the disease population enrolled, widespread application of results in clinical practice may not be warranted [10, 11]. One means to validate conclusions and extend the results to a more general population is replication of studies at novel sites. While conducting a second clinical trial to address the same hypothesis may be infeasible and potentially unethical, an observational study may provide reliable results using a fraction of the resources. Observational studies have been shown to be an effective means to reproduce results initially documented via clinical trials [106, 107]. Successfully replicating a study requires elements to be in place at the new site as well as sufficient documentation to be provided in the publication of the original study. The new site must have sufficient data for the population in question and a means of interrogating these data. The published study must detail the phenotype of the disease population in a way that can be reproduced at novel sites. Any risk models or calculations must be clearly explained if they are to be replicated.

In this and the following chapter, I describe initial evaluations of two studies and of UNMC's infrastructure in terms of the ability to replicate what was published from other sites. I describe the local site and publication requirements, the selection of the study, and review my results at attempted replication.

Site Requirements

Data Volume

As described earlier, small sample size is the leading cause of early termination of clinical trials. For studies running to completion, small sample size limits statistical power and the conclusions which can be drawn from the study. A researcher may miss a novel effect or may over-analyze in an attempt to find significance, both resulting in publication of false findings. Data extraction from the entire population recorded in an EHR may mitigate this problem. Rather than attempting to enroll dozens or even hundreds of patients, researchers have a pool of thousands or hundreds of thousands of patients to identify eligible patients from. All eligible patients may be enrolled in an observational study. In the case of rare conditions or diseases, extending the study to an entire distributed research network may see a ten-fold increase in the number of patients. In this way, even for rare diseases, it is likely a sufficient sample size may be identified to aid in drawing statistically sound conclusions.

Diverse Population

While a sufficiently large sample size is critical for drawing meaningful conclusions, it is critical that the study population is representative of the target population. The availability of diverse data within an EHR as well as increased diversity across a distributed research network, make it possible to have a representative study population, similar to what is described in the study being replicated. If the site attempting to replicate clinical studies only has limited date from a small registry, an appropriate study population may not be available.

Data Variety

Beyond having a sufficiently large and diverse sample, replicating clinical research may require a diverse set of data domains. Having a single domain, such as diagnostic codes, is useful for identifying information about disease incidence or prevalence. However, having only this information precludes further analysis in terms of disease progression, susceptibility, or outcomes. For these types of studies, additional data domains may be required. From the EHR, diagnostic history, laboratory data, medication orders and dispensing history, demographic information, procedures, and family history are often available. Some studies may involve extra-EHR data elements such as socioeconomic status, location information, or patient-reported outcomes from surveys. In order to attempt to replicate a study, all data elements from the original study must be available at the independent site. Missing data elements or data domains will limit the extent to which researchers can replicate prior results.

Infrastructure to Interrogate Data

With a data repository in place, to effectively replicate a study, researchers need an efficient means of interrogating the data, locally, or across a distributed research network. If individual data elements must have custom ETLs produced and custom SQL queries written, there will be a great deal of duplicated effort across studies and costs for developer time and infrastructure building may exceed the benefits of attempting to replicate a previous study's findings. With a clinical data research warehouse (CRDW) established, containing the majority of common data elements for most studies, a query platform may put in place to make data available to researchers. Across PCORnet, SAS queries are shared for this purpose (see appendix A for local instructions for running federated queries). Dozens of academic medical centers and other institutions have deployed i2b2 for local and federated queries. With minimal addition, these platforms allow researchers for diverse clinical studies to rapidly query existing data. Having this

infrastructure in place allows observational or retrospective studies to be accomplished with a fraction of the resources of conducting a new clinical trial.

Publication Requirements

Phenotype Description

While having a local infrastructure with sufficient data volume, data variety, and a querying mechanism in place is essential to be able to replicate studies, results and methods must be published in a replicable manner. A clear description of inclusion and exclusion criteria composing a phenotype is critical. Results may vary unnecessarily from the original study if a clear phenotype is absent. It is insufficient to simply state all patients with a disease of interest are included. Describing how this disease was defined for the study is essential. For instance, a diabetic study may include patients based on ICD diagnostic codes alone, may require one or multiple hemoglobin A1C (HbA1C) measures within a certain time frame, may be based on medication history, or some combination of any of these criteria. Each of these criteria should be provided in a human readable and machine parsable and interoperable format, possibly in the supplemental material to a published study. Without these clear descriptions, a replication attempt may study a very different population relative to what was originally documented.

Transparent Statistical Analysis

Finally, with the data and infrastructure in place and a study publishing a clear description of inclusion and exclusion criteria, it is also critical that the published work clearly document the statistical analysis and computations performed. Researchers have a vast arsenal of statistical means at their disposal, each of which may be appropriate in some situations. Understanding the original hypothesis, significance level, and any

model computation is requisite for a researcher attempting to replicate the study. A risk model whose components remain in a black box may not be replicable. If coefficients and how they were obtained is clear, a similar approach to developing a replicate is possible. If a winding path of statistical analysis with multiple endpoints and subsets of the population were employed and not published, replication may be infeasiable,

Study Selection

With the criteria described above, I evaluated a series of published risk models to attempt to replicate. This chapter focuses on efforts with one of these studies. While UNMC has a CRDW with an i2b2 infrastructure for querying a diverse set of clinical data, some studies required extra-EHR data or published insufficient information to enable an attempt at replication [109].

This initial evaluation centered on a study of 30-day hospital readmission for heart failure patients [109]. Readmission to a hospital within 30 days of discharge is a pervasive problem. With 19.6% of Medicare beneficiaries being readmitted to a hospital within 30 days of discharge the cost is estimated to be between $17.4 and $26 billion per annum [110, 111]. Section 3025 of the Affordable Care Act established the Hospital Readmissions Reduction Program (HRRP), which financially penalizes hospitals with excess readmissions (subpart 1 of 42 CFR part 412 (§412.150 through §412.154). Effective strategies to reduce unplanned readmissions are complex, resource intensive, and often short-lived [111, 112].

Identifying patients prior to discharge at the greatest risk for readmission may focus finite resources. Heart failure (HF) represents the primary cause of hospitalizations in patients over age 65 and is among the leading causes for preventable readmissions within 30 days [113, 114]. Understanding if the results published by this study are valid and

extendable to the population at UNMC is a necessary precursor to implementing any future change in clinical workflow or policy.

A number of models have been proposed and used to predict risk for 30-day readmission for heart-failure patients. For this assessment of coverage of UNMC's CRDW I selected Amarasingham's 2010 model for a number of reasons. First, this model encompasses much of preceding models [115]. Next, Amarasingham et al demonstrated the validity of their model in comparison to a CMS model and the Acute Decompensated Heart Failure Registry (ADHERE) model [109]. Finally, Amarasingham et al designed this model to utilize data commonly available in a basic EHR within the first 24 hours of hospital admission, making it likely UNMC's CRDW would contain much of the needed information. Additionally, this may facilitate future efforts of real-time application of the model within the EHR.

My initial hypothesis was the infrastructure established around UNMC's CRDW is sufficient to replicate a study that developed a risk model for readmission in heart failure patients. To address this hypothesis, I surveyed the necessary data elements from the study and produced the necessary queries to obtain sufficient data on the heart failure population at UNMC.

Methods –

Infrastructure

Overview

The clinical data research warehouse (CRDW) at the University of Nebraska Medical Center (UNMC) is a composite of de-identified data from a variety of sources (Figure 1). The development and use of this registry of patient data was approved by the institutional review board (IRB) at UNMC (IRB # 132-14-EP). Extracted patient-centric

data are transformed to adhere to vocabulary standards recommended by the office of the national coordinator (ONC) and loaded into informatics for integrating biology and the bedside (i2b2). Data are further transformed to conform to the National Patient-Centered Clinical Research Network (PCORnet) common data model (CDM) to allow participation in this national network [96, 97] .

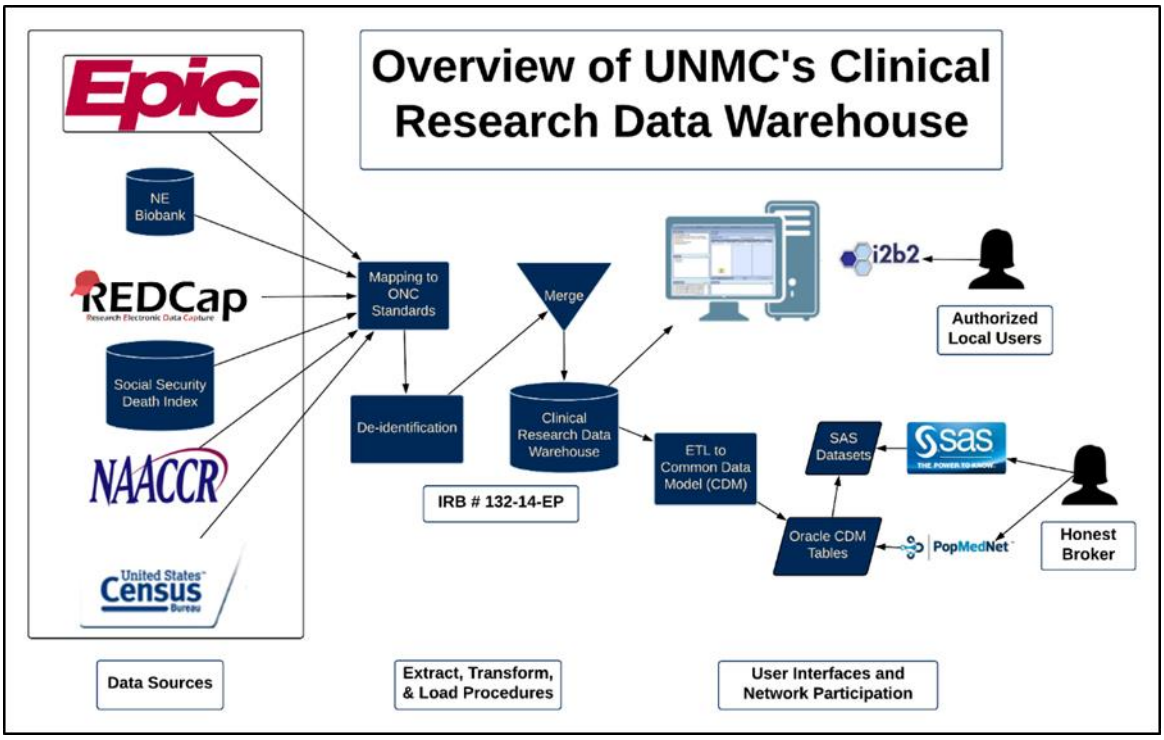

Figure 1. Overview of UNMC's CRDW. Data originate from both local and national sources. These data are extracted, transformed to ONC recommended vocabulary standards, and merged into a single data warehouse queryable by i2b2 and SAS.

Data Sources

Data for the clinical research are extracted from enterprise specific as well as national data sources. The bulk of data originate in Epic One Chart (Epic Systems Corporation, 1979 Milky Way, Verona, WI 53593), the electronic health record (EHR) at Nebraska Medicine. Local tissue biobank data and information from clinical research stored in Research Electronic Data Capture (REDCap) may also be incorporated to provide a more complete patient picture. These data are supplemented with information from the national level, including data from the Social Security Death Index (SSDI www.ntis.gov/products/ssa-dmf.aspx), the North American Association of Central Cancer Registries (NAACCR www.naaccr.org), and from the United States Census (www.census.gov).

UNMC's CRDW has de-identified information for more than 2 million patients. The majority of this data were recorded since 2012 when Nebraska Medicine adopted an Epic EHR. However, some data, especially demographic and laboratory information, has coverage originating many years earlier. For more than half a million patients, relatively complete medical information exists for the past several years in this database.

De-Identification:

All data stored in UNMC's CRDW is fully de-identified according to guidelines associated with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule [116] . Protected health information (PHI), including patient names, addresses, e-mail addresses, phone numbers, social security numbers (SSN), fax numbers, health insurance beneficiary numbers, account numbers, medical record numbers, certificate/license numbers, vehicle identification numbers, and other unique identifying numbers are excluded from the data warehouse. All dates, including birth dates,

encounter dates, death dates, and diagnoses or procedure dates, are obfuscated. On a per patient basis, dates are shifted by a random number from -1 to -30. Date shifting is consistent across all of a single patient's encounters and characteristics. Geographical identifiers are obfuscated to ensure the covered population is greater than 20,000 individuals. To accomplish this, zip codes are truncated after two or three digits.

Standardization

As data are extracted from disparate and often proprietary formats, we dedicated considerable effort to transform elements to ONC recommended vocabulary standards [77, 117] . Where feasible, standards were adopted or recommended to facilitate interoperability. Per ONC guidelines, encounter diagnoses were mapped to international classification of disease (ICD) versions 9 and 10, demographic data such as race and ethnicity conform to Office of Management and Budget (OMB) standards, laboratory information are represented via Logical Observation Identifiers Names and Codes (LOINC www.loinc.org), medication information is presented with appropriate National Drug Codes (NDC) and RxNorm codes, and clinical findings are reported as Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) terms (See the Interoperability Standard Advisory from the Office of the National Coordinator for Health IT [118] for further details of datatypes and standards).

Informatics for Integrating Biology and the Bedside (i2b2)

The CRDW is built on the Informatics for Integrating Biology and the Bedside (i2b2) research environment. This is an NIH-funded National Center for Biomedical Computing (NCBC) devoted to translational research (http://www.i2b2.org). In particular, it is a scalable, open-source informatics framework and architecture that can be used to host a research data warehouse [119] . Developed within the Partner's Health Care, i2b2 is

now deployed at more than 60 academic medical centers across the country. This web-based service allows authorized users to query de-identified patient data for exploratory analysis and cohort identification. With a common messaging protocol, queries may be developed locally, tested on a single date warehouse, and then shared across large networks to interrogate data covering many more patients. In this way, the query, rather than patient data, is shared between institutions.

Model Replication

Defining the Study Population

Eligible patients (table 2) have a documented diagnosis of heart failure based on ICD 9 code [109, 120] and were admitted to a Nebraska Medicine hospital at least once from May 2012 to December 2015. Visits for rehabilitation (DRG 462) were excluded. Additionally, admissions where the patient expired or left against medical advice (ICD10 CM Z53.21, ICD9CM: V64.2) were not considered as index hospitalizations [110]. A 30-day readmission was defined as an admission to the hospital for any cause within 30 days of the most recent discharge from an acute care hospital.

Table 1 - Patient Cohort Identification		
Table 1A - Inclusion Criteria		
Hospital Admission Between May 2012 and December 2015 (CPT4 Codes)	99221, 99222, 99223	Initial Inpatient Hospital Care Procedures
	99231, 99232, 99233	Subsequent Inpatient Hospital Care Procedures
	99234, 99235, 99236	Observation or Inpatient Care Services (Including Admissions and Discharge Services)
	99238, 99239	Hospital Discharge Services
Diagnosis of Heart Failure (ICD9 CM Codes)	402.01	Malignant hypertensive heart disease with heart failure
	402.11	Benign hypertensive heart disease with heart failure
	402.91	Unspecified hypertensive heart disease with heart failure
	404.01	Hypertensive heart and kidney disease, malignant, with heart failure and with chronic kidney disease stage I through stage IV, or unspecified
	404.03	Hypertensive heart and chronic kidney disease, malignant, with heart failure and chronic kidney failure stage V or end stage renal disease
	404.11	Hypertensive heart and renal disease, benign, with heart failure
	404.13	Hypertensive heart and renal disease, benign, with heart failure and renal failure
	404.91	Hypertensive heart and renal disease, unspecified, with heart failure
	404.93	Hypertensive heart and renal disease, unspecified, with heart failure and renal failure
	425.1	Hypertrophic cardiomyopathy
	425.4	Other primary cardiomyopathies
	425.5	Alcoholic cardiomyopathy
	425.7	Nutritional and metabolic cardiomyopathy
	425.8	Cardiomyopathy in other diseases classified elsewhere
	425.9	Secondary cardiomyopathy, unspecified
	428.xx	Heart Failure

Table 1B - Exclusion Criteria
Hospitalizations when patient left against medical advice (ICD10CM - Z53.21)
Admissions for rehabilitation (DRG 462)
Exclude index when patient expired

Table 2. Eligibility Criteria for Inclusion in the Amarasingham Model. All patients were admitted to a Nebraska Medicine Hospital between May 2012 and December 2015.

Assessing Coverage

For assessing the coverage of data, we considered all inpatient encounters for eligible patients during the date range specified. Patient counts were generated using i2b2 queries.

Variable Definitions

Age was defined by the age at the initial visit. Laboratory values were defined by relevant LOINC codes. These were not defined by the original authors. History of depression or anxiety was defined by the following ICD-9-CD codes: 293.83, 293.84, 297.1, 297.9, 300.0, 300.00, 300.01, 300.02, 300.09, 300.11, 300.23, 300.29, 300.4, 300.9, 301.6, 301.83, 301.9, 306.1, 306.4, 308.0, 308.3, 308.9, 309.0, 309.24, 309.28, 309.29, 309.9, 311, and 312.20. Appendix B contains a printed version of a sample i2b2 query's generated SQL for identifying the variables from the original risk model.

Obtaining Missing Data Elements

Clinicians and laboratory technicians were consulted to determine if missing data elements were routinely ordered or recorded at Nebraska Medicine. Following the pattern for previous data extracts, data elements were identified in Clarity data tables, and extracted, transformed to align with ONC standards where applicable, and loaded into the star schema supporting i2b2. In addition to loading new facts, metadata was created as necessary to allow querying in the web client. As part of this metadata creation and updates, metadata xml was created for many existing labs in order to allow the user to query lab values rather than just the presence or absence of results. Finally, quality assessment on new and existing data was performed to ensure proper loading and mapping. This was done by sampling patients and encounters and reviewing pertinent records from Clarity.

Results

Data Coverage

The CRDW at UNMC has excellent overall coverage of the variables in the Amarasingham risk model (34/40, 85%) (Table 3). All variables in the demographics and laboratory values are covered. Vital signs will be fully covered if the Tabak mortality score is calculated and recorded. Socioeconomic status and health system interactions had the poorest coverage (50% and 66%, respectively).

Table 2 - Model Coverage in CRWD		
Variable		**Present in CRDW (n = 4,915 patients)**
Laboratory Values		
Albumin (g/dL)*		4,245 patients
CPK*		2,665 patients
Creatinine (mg/dL)*		4,858 patients
Na (mEq/dL)*		4,858 patients
BUN (mg/dL)*		4,858 patients
Arterial pH*		2,430 patients
Arterial pCO2 (mm Hg)*		2,430 patients
Troponin I or CKMB*		3,097 patients
PT INR*		3,854 patients
Total bilirubin (mg/dL)*		4,243 patients
WBC (k/mm^3)*		4,796 patients
Glucose*		4,858 patients
BNP (b)*		2,848 patients
proBNP (b)*		0 patients
Demographics		
Age*	>45	3,742 patients
	<=45	362 patients
Gender	Male	2,702 patients
	Female	2,213 patients
Race	Black	637 patients
	Non-black	4,304 patients
Ethnicity	Hispanic	0 patients
	Non-hispanic	4,915 patients
Vital signs		
Temperature (F)		4,892 patients
Pulse (per minute)*		4,893 patients
Systolic BP (mm Hg)*		4,891 patients
Diastolic BP (mm Hg)*		4,891 patients
Altered mental status*		4,887 patients
Tabak Score		May be calculated
Diagnoses (ICD-9 based)		
COPD*		1,742 patients
Chronic Pulmonary Heart Disease*		922 patients
Metastatic cancer*		66 patients
History of Depression or Anxiety (past 12 months)		961 patients
Socioeconomic Status		
Living Alone (Single, Widowed, or Divorced at index)		4,915 patients
Residence census tract in lowest socioeconomic quintile^		Need link to census
# of home address changes prior to index		Availabe within EHR
Payer information (insurance type)^^		Availabe within EHR
History of cocaine use (past 12 months)		19 patients
Left against medical advice (past 12 months)		9 patients
Health System Interactions		
Missed scheduled visit (past 12 month period)		Availabe within EHR
Uses health system pharmacy (past 12 months)		4,914 patients
# of prior inpatient admissions (past 12 months)		4,915 patients
# of ED visits (past 12 months)		3,052 patients
# of Outpatient visits (past 12 months)		3,915 patients
Presented to ED between 6:00 a.m. to 6:00 p.m. (index)		De-identification

*Tabak 2007 Model;
^Based on census data: Median household income, percent of households above poverty, percent of residents with age >= 18 with college/higher ed, percent of pop. who were white;
^^Medicare, Medicaid, Commercial insuracnce, Self-pay, Other, Uninsured

Table 3. Amarasingham Model Data Coverage in UNMC's CRDW. The majority (34/40) variables were represented in UNMC's CRDW.

Discussion

Conclusions

I rejected my hypothesis that the current CRDW at UNMC was sufficient to fully replicate this risk model for readmission in heart failure patients. Inability to replicate the study stemmed both from insufficient information being published and from lack of sufficient data variety being available and incorporated into the CRDW.

First, the authors failed to publish how the risk model coefficients were calculated for application of their scoring. Attempts to reach the author for clarification went unanswered. In addition, standard LOINC codes for laboratory values were not published necessitating the replication effort to include attempting to identify appropriate laboratory tests. Without these data, full replication is impossible.

In addition to insufficiencies in the published study, limitations in data variety in the CRDW prevented full replication of the risk model. Socioeconomic status required linking EHR data to U.S. Census data which had not been completed at UNMC. Multiple variables, including number of home address changes, payer information, and missed scheduled visits may be extracted from the EHR in the future. Finally, arrival time at the ED was excluded as this work was completed in a de-identified database.

While data elements present in the de-identified CRDW did not allow for full replication of this risk model, this assessment helps prioritize efforts to load additional clinical data elements into the CRDW.

Limitations

Data for this study come only from Nebraska Medicine and do not reflect readmissions to other hospitals. Nasir et al indicate same-hospital readmission rate is not a reliable indicator of all-hospital readmission rates. They showed only 80.9% of

readmissions for heart failure patients were to the index hospital [85] . However, same-hospital data may be a useful benchmark for internal quality improvement. Readmission data for all hospitals in the area may be available through NeHII and may be used in future studies.

Future Research

Future research may focus on validating the existing model and investigating new variables to include. Additionally, the queries used in this research effort may be shared with other sites using i2b2 to expand the study cohort. Further research can be formed with similar methods for other patient cohorts who are commonly readmitted (COPD, Pneumonia, AMI, TKA/THA, and septicemia patients). Eventually, this predictive model may be applied in real-time to identify patients at risk for readmission before they leave the hospital or as they present to the ED.

CHAPTER 2 – ATTEMPTED REPLICATION OF AN IN-HOSPITAL MORTALITY RISK MODEL

Introduction

As the heart failure study proved irreplicable, I selected another risk model addressing in-hospital mortality. Tabak's 2014 model for mortality risk has demonstrated validity relative to published models and was designed to use data commonly available in an EHR in the first 24 hours of a patient visit [121]. This risk model is based on age, gender, and 23 laboratory findings.

Methods

Infrastructure

See chapter one for a full description of the development and contents of the UNMC CRDW. Data coverage and encounter data were collected via i2b2 queries and SQL scripts targeting the de-identified database (Appendix C).

Assessing Coverage

For assessing the coverage of our current data, we considered all encounters with any discharge disposition recorded (Logical Observation Identifiers Names and Codes (LOINC) 75528-0 and 75527-2) for a three-year period (de-identified 5/1/2012 – 5/1/2015). Patient counts to assess data coverage were generated using i2b2 queries using pertinent LOINC and Current Procedure Terminology, 4th Version (CPT-4) codes.

Defining the Study Population

Data were analyzed for patients discharged from a Nebraska Medicine hospital between 1/1/2013 and 12/31/2017 (de-identified date range). Patients with a discharge were defined by having a discharge disposition recorded based on LOINC code 75528-0 and a vital status recorded at discharge based on LOINC code 75527-2. Those patients with missing data or a status of unknown, no information, or other recorded were excluded. Only inpatient encounters were considered. Encounters with a length of stay less than 24 hours were excluded.

Outcome variable

The outcome variable for this study was inpatient mortality. Inpatient mortality was defined by a vital status at discharge of “Expired” (LOINC 75527-2 code “E”) or a discharge disposition of expired (LOINC 75528-0). Encounters with unknown, other, or no information recorded for vital status at discharge were excluded.

Statistical Analysis

Using the coefficients provided by Tabak et al, a risk score was calculated for all encounters in the cohort. This score was used as the predictor variable with inpatient mortality as the dependent variable to fit a logistic regression model.

Missing Values

For missing lab values, we followed Tabak et al’s procedure of assigning a score based on the reference range for the missing variable [121]. For each lab, if multiple values exist, the earliest recorded value in the encounter is used.

Creating a Receiver Operator Characteristic (ROC) curve and computing the area under the curve (AUC)

With missing values imputed as described above, a risk score was calculated for each encounter as the sum of scores for each data element. For each risk score, the number of expired and living patients at discharge was identified and the sensitivity and specificity for each score as a cutoff was computed. Next, for each risk score the true positive rate (sensitivity) and the false positive rate (1 – specificity) were tabulated. Based on this table, an ROC curve was plotted and the area under the curve was calculated (concordance (c)-statistic). Analyses performed with R (R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.).

Logistic Regression and Comparing c-statistics

Using functions available in R packages (glm) we fit a logistic regression to our data. We compared the coefficients generated from these data to those reported by Tabak et al. To assess the predictive power of our model, we computed the c-statistic for the logistic regression. We then compared this c-statistic to that published by Tabak et al [121] . A c-statistic of 0.5 indicates a predictive model no better than chance. As the c-statistic increases toward 1.0, the discriminatory power of the model is greater. The c-statistic defines the probability that a patient selected from the outcome group will have a risk score greater than a patient selected from the group without the outcome of interest.

Results

Data Coverage

All variables employed by the Tabak risk model were present in the CRDW at UNMC (25/25, 100%) (see table 4). The Demographics variables had the best coverage, with over 99% of encounters having this data available. Serum chemistry variables

demonstrated good coverage with between 57% and 80% of all encounters having recorded values. Arterial blood gas labs and cardiac markers demonstrated the lowest coverage. Less than 16% of encounters had arterial blood gas values. Three of the six hematology and coagulation variables were represented in over 80% of the encounters while the other three variables had less than 40% of encounters with results (see table 4).

Encounter Characteristics

For the specified date range, 79,039 distinct adult patients were identified having a total of 136,084 distinct, inpatient encounters lasting longer than 24 hours. The number and percentage of these encounters with each of the risk score variables is displayed in table **5. For these encounters, 3,047 (2.24%) resulted in a discharge disposition of expired. The average total score for encounters where the patient expired was 77.0 (SD 23.0) while the average total score for encounters where the patient was discharged alive was 34.8 (SD 21.0) (Figure 2).

Variable	# of Encounters (n = 136,084 (%))
Demographics	
Age >= 18	136,084 (100%)
Gender	136,084 (100%)
Comprehensive Metabolic Panel	
Potassium	116,057 (85.3%)
Sodium	115,972 (85.2%)
Calcium	116,097 (85.3%)
Creatinine	115,974 (85.2%)
Glucose	111,710 (82.1%)
BUN	115,828 (85.1%)
Albumin	85,424 (62.8%)
AST	84,443 (62.1%)
Total Bilirubin	84,984 (62.4%)
Alkaline Phosphatase	84,423 (62%)
Arterial Blood Gas (ABG)	
pH Arterial	21,624 (15.9%)
PO2 Arterial	20,738 (15.2%)
PCO2 Arterial	21,625 (15.9%)
Complete Blood Count	
Bands	26,613 (19.6%)
Hemoglobin	126,950 (93.3%)
Platelets	122,773 (90.2%)
WBC	122,428 (90%)
Cardiac Biomarkers	
Pro-BNP	133 (0.1%)
BNP	16,687 (12.3%)
PTT	35,202 (25.9%)
PT-INR	58,958 (43.3%)
Troponin I or CPK MB	37,194 (27.3%)

Table 4. Number of Encounters Recording Variables from the Tabak Mortality Risk Model

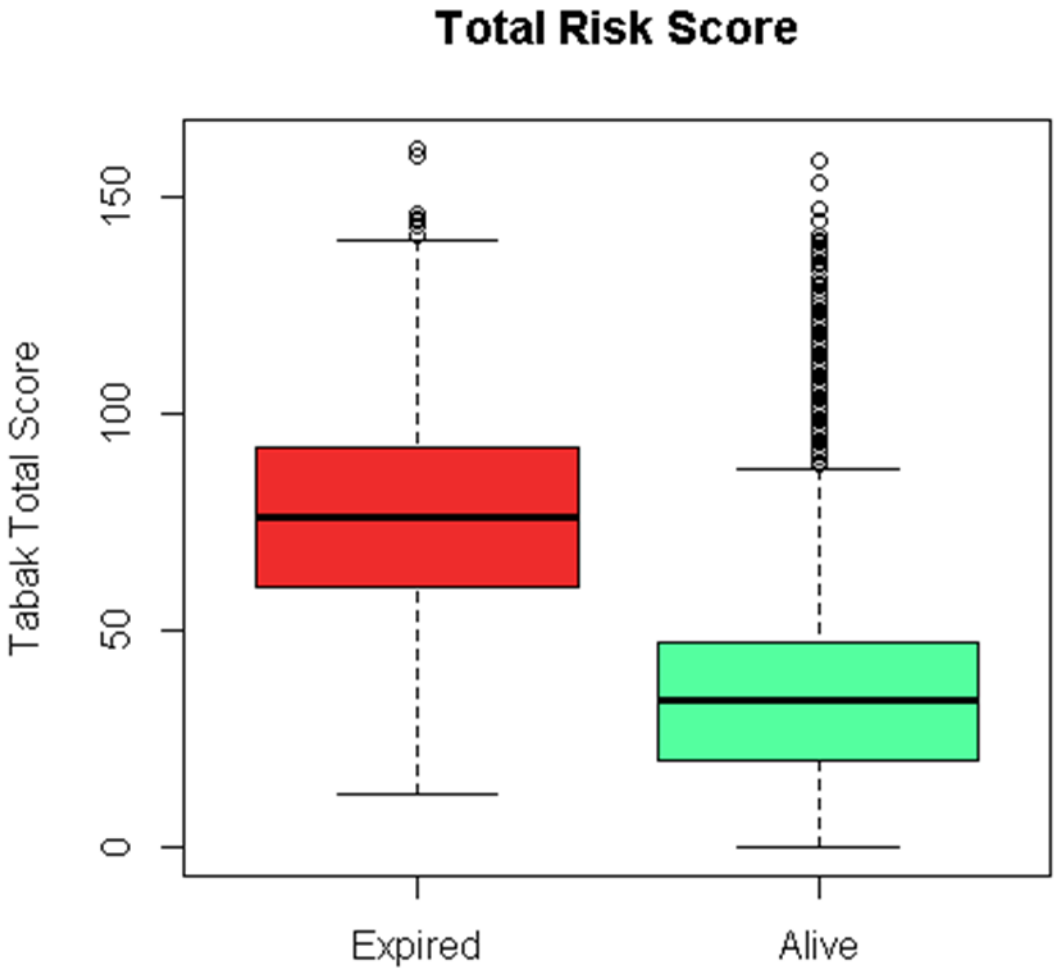

Figure 2. Comparison of Tabak Total Score for Patients Discharged Relative to Expired Patients.

Receiver Operator Curve

Using the total score with Tabak's reported coefficients as the predictor, a receiver operator characteristic curve was computed (Figure 3a). As it is up and to the left, the curve demonstrates the high predictive value of the model. The area under the curve (AUC) is 0.94. As the dataset is highly unbalanced, we also created a precision-recall (PR) curve (Figure 3b).

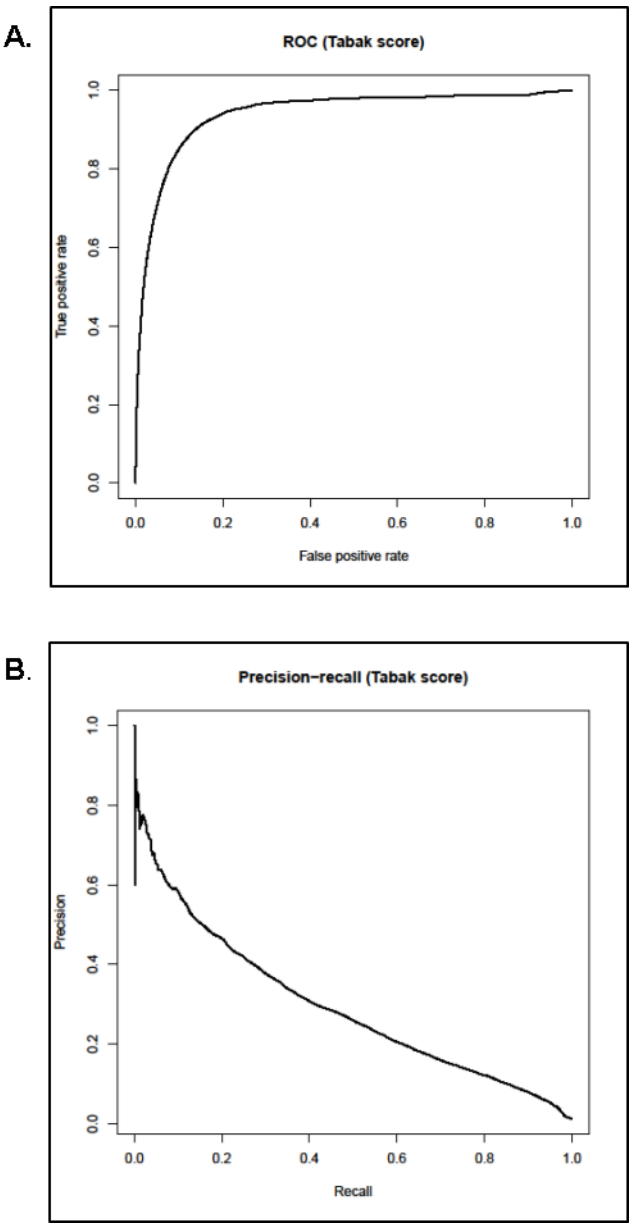

Figure 3. Receiver Operator Characteristic (ROC) curve and Precision-Recall (PR) Curve based on Tabak mortality score. ROC (A) and PR (B) Curves based on scores calculated from 136,084 inpatient encounters for adults at Nebraska Medicine between January 2013 and December 2017. Area under the ROC curve (AUC) is 0.94.

Logistic Regression

Using the ranges assigned by Tabak et al for age, gender, and each laboratory test, we used multiple logistic regression to compute new coefficients for a risk model score (Table 5). Recomputed coefficients demonstrated medium correlation to the coefficients reported in the original study (Spearman rho = 0.709, $p < 0.001$) (Figure 4).

Variable	Discharges, n (%)	Mortality, n (%)	Coeff.	Std. Error	P	Sig.	Tabak Coeff.	Diff.
30-34	9,026 (6.6)	53 (0.6)	0.24	0.19	0.212		0.21	0.03
35-39	7,316 (5.4)	56 (0.8)	0.38	0.19	0.046	*	0.67	-0.29
40-44	7,117 (5.2)	70 (1)	0.34	0.18	0.057	.	0.83	-0.49
45-49	8,284 (6.1)	134 (1.6)	0.68	0.16	< 0.001	***	1.12	-0.44
50-54	11,197 (8.2)	198 (1.8)	0.79	0.15	< 0.001	***	1.28	-0.49
55-59	13,545 (10)	286 (2.2)	0.89	0.14	< 0.001	***	1.47	-0.58
60-64	13,857 (10.2)	374 (2.8)	1.14	0.14	< 0.001	***	1.64	-0.50
65-69	12,970 (9.5)	387 (3.1)	1.24	0.14	< 0.001	***	1.8	-0.56
70-74	10,600 (7.8)	360 (3.5)	1.44	0.14	< 0.001	***	1.96	-0.52
75-79	8,946 (6.6)	308 (3.6)	1.57	0.14	< 0.001	***	2.11	-0.54
80-84	7,403 (5.4)	313 (4.4)	1.81	0.14	< 0.001	***	2.32	-0.51
85-89	7,624 (5.6)	317 (4.3)	1.93	0.14	< 0.001	***	2.51	-0.58
>89	448 (0.3)	113 (33.7)	3.77	0.18	< 0.001	***	2.78	0.99
Male	61,124 (44.9)	1,673 (2.8)	0.12	0.04	0.007	**	0.14	-0.02
Albumin <= 2.4 g/dL 4	7,222 (5.3)	795 (12.4)	1.17	0.07	< 0.001	***	0.89	0.28
Albumin 2.5 - 2.7 g/dL	6,298 (4.6)	374 (6.3)	0.72	0.08	< 0.001	***	0.47	0.25
Albumin 2.8 - 3 g/dL	9,305 (6.8)	453 (5.1)	0.63	0.07	< 0.001	***	0.26	0.37
Albumin 3.1 - 3.3 g/dL	12,842 (9.4)	431 (3.5)	0.44	0.07	< 0.001	***	0.08	0.36
AST 31 - 40 U/L	8,997 (6.6)	349 (4)	0.31	0.07	< 0.001	***	0.14	0.17
AST 41 - 60 U/L	7,401 (5.4)	397 (5.7)	0.46	0.07	< 0.001	***	0.28	0.18
AST 61 - 100 U/L	5,149 (3.8)	310 (6.4)	0.41	0.08	< 0.001	***	0.37	0.04
AST > 100 U/L	6,553 (4.8)	578 (9.7)	0.60	0.07	< 0.001	***	0.6	0.00
Total Bilirubin 1.5 - 2 mg/dL	4,552 (3.3)	249 (5.8)	0.25	0.08	0.003	**	0.07	0.18
Total Bilirubin > 2.0 mg/dL	6,414 (4.7)	571 (9.8)	0.36	0.07	< 0.001	***	0.29	0.07
Calcium <= 7.9 mg/dL L	7,523 (5.5)	590 (8.5)	-0.03	0.07	0.635		0.26	-0.29
Calcium 8 - 8.4 mg/dL	16,221 (11.9)	583 (3.7)	-0.08	0.06	0.145		0.09	-0.17
Calcium > 10.1 mg/dL	3,727 (2.7)	132 (3.7)	0.40	0.11	< 0.001	***	0.22	0.18
Creatinine > 2.0 mg/dL	13,715 (10.1)	800 (6.2)	-0.12	0.07	0.072	.	0.09	-0.21
pro BNP 8001 - 18000	12 (0)	2 (20)	2.16	0.83	0.009	**	0.35	1.81

pro BNP > 18000	9 (0)	0 (0)	-9.18	99.82	0.927		0.68	-9.86
BNP 1201 - 2400	1,594 (1.2)	144 (9.9)	0.21	0.11	0.057	.	0.11	0.10
BNP > 240	733 (0.5)	89 (13.8)	0.58	0.14	< 0.001	***	0.28	0.30
Glucose <= 70 mg/dL	1,754 (1.3)	113 (6.9)	0.45	0.13	< 0.001	***	0.43	0.02
Glucose 136 - 165 mg/dL	16,248 (11.9)	525 (3.3)	0.26	0.06	< 0.001	***	0.16	0.10
Glucose > 165 mg/dL	19,837 (14.6)	781 (4.1)	0.28	0.05	< 0.001	***	0.32	-0.04
K <= 3.2 mEq/L	8,065 (5.9)	255 (3.3)	0.09	0.08	0.242		0.19	-0.10
K 5 - 5.3 mEq/L	3,812 (2.8)	212 (5.9)	0.05	0.09	0.572		0.1	-0.05
K > 5.3 mEq/L	3,564 (2.6)	287 (8.8)	0.11	0.08	0.206		0.21	-0.10
Na <= 130	8,503 (6.2)	445 (5.5)	0.12	0.07	0.066	.	0.28	-0.16
Na 131 - 135	27,553 (20.2)	836 (3.1)	0.07	0.05	0.162		0.09	-0.02
Na 144 - 145	1,876 (1.4)	96 (5.4)	0.40	0.13	0.002	**	0.27	0.13
Na > 145	995 (0.7)	114 (12.9)	0.69	0.13	< 0.001	***	0.61	0.08
Alk Phos 116 - 220 U/L	14,430 (10.6)	576 (4.2)	0.01	0.06	0.898		0.11	-0.10
Alk Phos 221 - 630 U/L	4,109 (3)	238 (6.1)	0.04	0.09	0.675		0.34	-0.30
Alk Phos > 630 U/L	636 (0.5)	59 (10.2)	0.31	0.17	0.068	.	0.54	-0.23
BUN 26 - 30 mg/dL	7,032 (5.2)	319 (4.8)	0.33	0.07	< 0.001	***	0.24	0.09
BUN 31 - 40 mg/dL	7,783 (5.7)	386 (5.2)	0.15	0.07	0.035	*	0.37	-0.22
BUN 41 - 55 mg/dL	5,692 (4.2)	344 (6.4)	0.23	0.08	0.005	**	0.53	-0.30
BUN > 55 mg/dL	5,630 (4.1)	452 (8.7)	0.36	0.09	< 0.001	***	0.68	-0.32
pH Arterial <= 7.2	1,401 (1)	384 (37.8)	1.72	0.09	< 0.001	***	1.38	0.34
pH Arterial 7.21 - 7.3	2,942 (2.2)	466 (18.8)	1.27	0.07	< 0.001	***	0.87	0.40
pH Arterial 7.31 - 7.35	3,355 (2.5)	376 (12.6)	1.11	0.07	< 0.001	***	0.66	0.45
pH Arterial > 7.48	1,497 (1.1)	180 (13.7)	0.58	0.10	< 0.001	***	0.5	0.08
PO2 <= 50	479 (0.4)	81 (20.4)	0.90	0.15	< 0.001	***	0.79	0.11
PO2 50.1 - 55	486 (0.4)	81 (20)	0.91	0.15	< 0.001	***	0.57	0.34
PO2 > 140	7,781 (5.7)	639 (8.9)	0.34	0.06	< 0.001	***	0.78	-0.44
pCO2 Arterial <= 35	6,758 (5)	881 (15)	0.97	0.06	< 0.001	***	0.56	0.41
pCO2 Arterial >50	3,513 (2.6)	533 (17.9)	0.86	0.08	< 0.001	***	0.46	0.40
PTT <= 22	653 (0.5)	43 (7)	0.72	0.19	< 0.001	***	0.22	0.50

PTT 45.1 - 55	1,324 (1)	148 (12.6)	0.33	0.11	0.003	**	0.21	0.12
PTT > 55	1,532 (1.1)	225 (17.2)	0.61	0.10	< 0.001	***	0.28	0.33
PT INR 1.11 - 1.4	12,828 (9.4)	664 (5.5)	0.22	0.06	< 0.001	***	0.23	-0.01
PT INR 1.41 - 2	6,000 (4.4)	471 (8.5)	0.41	0.07	< 0.001	***	0.44	-0.03
PT INR 2.1 - 5	5,942 (4.4)	456 (8.3)	0.68	0.07	< 0.001	***	0.34	0.34
PT INR > 5	765 (0.6)	81 (11.8)	0.68	0.15	< 0.001	***	0.51	0.17
Bands 7 - 13%	5,126 (3.8)	451 (9.6)	0.85	0.07	< 0.001	***	0.37	0.48
Bands 14 -32%	6,641 (4.9)	644 (10.7)	0.80	0.06	< 0.001	***	0.59	0.21
Bands > 32%	2,565 (1.9)	309 (13.7)	0.87	0.08	< 0.001	***	0.79	0.08
Hemoglobin <= 10 g/dL	24,558 (18)	1,107 (4.7)	0.20	0.05	< 0.001	***	0.16	0.04
Hemoglobin > 18 g/dL	477 (0.4)	24 (5.3)	0.18	0.25	0.474		0.25	-0.07
Platelets <= 115*109/L	10,019 (7.4)	667 (7.1)	0.34	0.06	< 0.001	***	0.63	-0.29
Platelets 115.1 – 150*109/L	11,084 (8.1)	352 (3.3)	0.07	0.07	0.284		0.13	-0.06
Platelets > 420*109/L	5,249 (3.9)	131 (2.6)	-0.19	0.11	0.077	.	0.13	-0.32
WBC <= 4.3*1,000/mm3	25,421 (18.7)	480 (1.9)	-0.11	0.06	0.059	.	0.27	-0.38
WBC 11 - 14.1*1,000/mm 3	14,674 (10.8)	453 (3.2)	0.30	0.06	< 0.001	***	0.28	0.02
WBC 14.2 - 19.8 *1,000/mm3	9,182 (6.7)	417 (4.8)	0.33	0.07	< 0.001	***	0.47	-0.14
WBC > 19.8*1,000/mm 3	3,380 (2.5)	282 (9.1)	0.32	0.08	< 0.001	***	0.78	-0.46
Troponin I 0.05-0.1 or CPK MB 3-5 ng/mL	5,542 (4.1)	435 (8.5)	0.62	0.06	< 0.001	***	0.15	0.47
Troponin I 0.11-0.2 or CPK MB 6-10 ng/mL	1,709 (1.3)	193 (12.7)	0.86	0.10	< 0.001	***	0.29	0.57
Troponin I 0.21-0.3 or CPK MB 11-34 ng/mL	744 (0.5)	99 (15.3)	0.94	0.14	< 0.001	***	0.54	0.40
Troponin I >0.3 or CPK MB >34 ng/mL	2,810 (2.1)	392 (16.2)	0.98	0.07	< 0.001	***	0.82	0.16
Total	136,084 (100)	3,047 (2.3)						

Table 5. Logistic Regression Coefficients for Variables in Tabak Mortality Score. AST = Aspartate Aminotransferase, Pro-BNP = pro-B-type natriuretic peptide, BNP = B-type

natriuretic peptide, K = Potassium, Na = Sodium, Alk Phos = Alkaline Phosphatase, BUN Blood Urea Nitrogen, PTT = Prothrombin Time, Prothrombin Time International Normalized Ratio, WBC = White Blood Cells, CPK MB = Creatinine Kinase Muscle Brain.

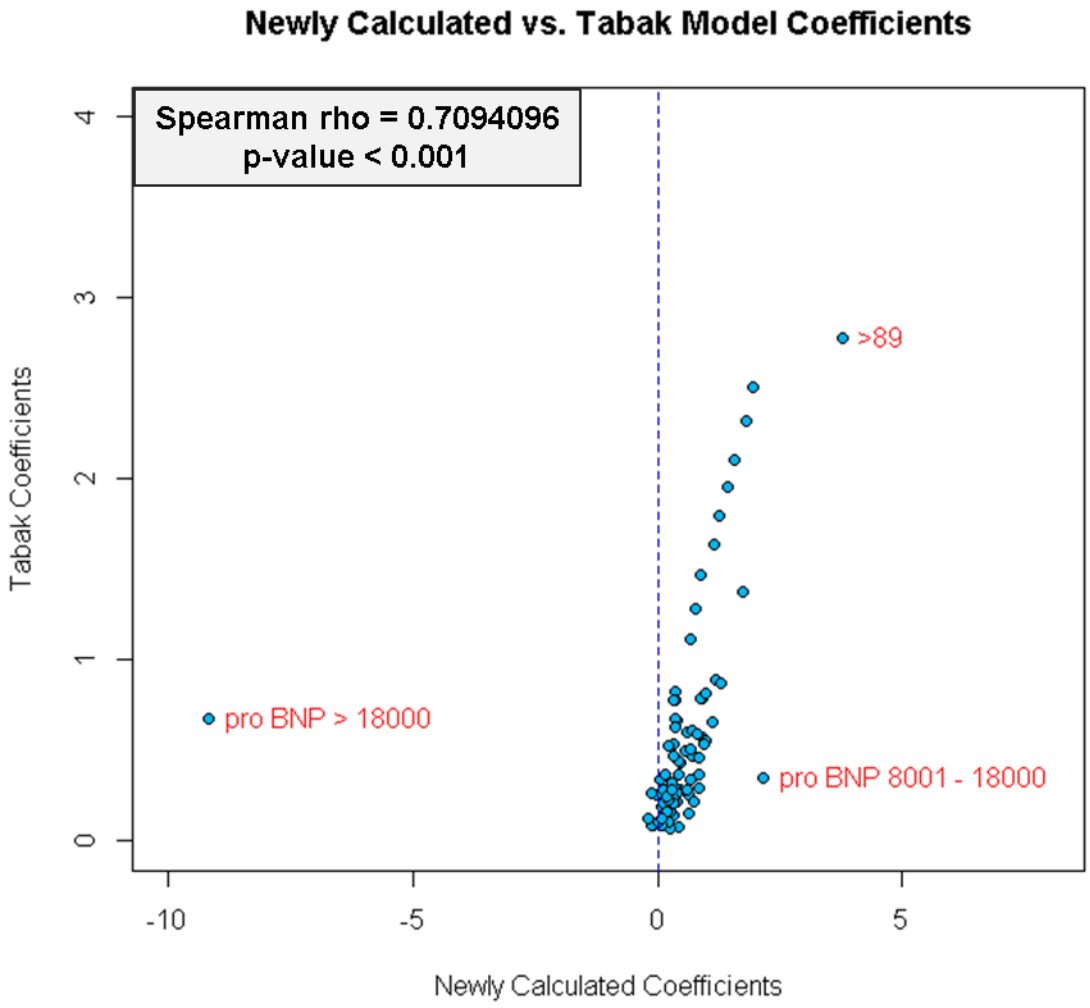

Figure 4. Correlation of Newly Calculated and Tabak Model Coefficients. Labeled points had a difference with magnitude greater than 0.75.

Discussion

Conclusions

Sufficient data volume and variety were included in UNMC's CRDW to enable replication of the readmission risk model. Replication was further facilitated as the authors published clear methods on how they computed the coefficients for the risk model. In addition, the original article clearly defined the characteristics of the encounters to be included in the analysis. With the infrastructure in place and sufficient details published, we successfully replicated this risk model and extended the results to a novel population. This lends further credence to the original work and supports use of the model for further studies or applications within CDS.

A few points comparing newly calculated coefficients to the originally published values stand out. The Pro-BNP values demonstrate a large discrepancy between this replication effort and the original study. This is most likely due to the very low encounter counts recording this laboratory value (21 total encounters with abnormal values recorded). The apparent protective effect of a grossly elevated pro-BNP is most likely an artifact of zero of the nine encounters resulting in an expired patient. Further analysis of this model may benefit from excluding this variable as it has such a low encounter count. The coefficient for age > 89 being slightly different from the original study is likely also due to a low encounter count for this category coupled with a slightly higher incidence of mortality for this category. Further study may require a threshold for a minimum encounter count for a variable to be included in the analysis.

Limitations

This study was limited as data were obtained solely from Nebraska Medicine affiliated hospitals. The study being replicated collected data from 70 hospitals and

accumulated nearly ten times the number of encounters to include. Artifacts due to the local environment and workflows may introduce bias in our analysis. In addition, the demographics included in our study sample may not be representative of the entire United States and may not be applicable in all areas.

Future Research

With successful replication of this readmission risk model, applying this replication methodology to additional risk models is possible. The pattern of evaluating the publication and assessing the coverage of the CRDW infrastructure provides a means to potentially replicate any clinical research.

In addition, with the validation and extension of the risk model described in this chapter, further analyses on subsets of the population are possible. Refining the risk model for specific diseases as well as considering additional variables with high predictive power are possible.

Finally, incorporating this risk model score into the CRDW and eventually into CDS is possible. Within i2b2, the score can be calculated and stored on a per encounter basis. Metadata may be developed to allow researchers to use this score within clinical queries. As the efficacy of the risk model are further demonstrated, incorporating this risk score upstream within the EHR becomes possible.

CHAPTER 3 – DEVELOPING A COMPUTABLE PHENOTYPE FOR PREGNANCY

Introduction

Distributed research networks relying on data extracted from the EHR offer a variety of benefits for comparative effectiveness research (CER) and conducting pragmatic trials. Data are collected for millions of patients including diverse demographics and clinical environments and are representative of actual care [97-99]. Federated queries allow patient screening and subsequent data collection with a fraction of previously required resources. However, to be effective, queries must be applicable across all sites. Each query represents a computable phenotype, i.e. a set of characteristics and clinical features more commonly observed in patients with a disease or condition than individuals in the general population [122]. Some elements of these phenotypes recur as building blocks for queries in other studies. For instance, many queries seek to identify a population with diabetes. Some of these may focus on individuals of a certain age, only inpatients, or only those with specified comorbidities. In each instance, the base population of those with diabetes must be defined. For many diseases, professional societies have put forth detailed guidelines that can be adopted to construct interoperable computable phenotypes.

One common condition for which a computable phenotype has not been well-defined is pregnancy. While the EHR is replete with data on this condition and being pregnant is very often an inclusion or exclusion criteria for clinical research, no standard EHR definition has been put forth to consistently identify a pregnant population [123]. Defining this element has often been done ad hoc with limited success [124]. Some

challenges to creating a common definition include the variety of measures and findings that may indicate a pregnancy, the diverse location in the EHR these elements may be stored, inconsistent coding across sites, and the varying levels of care sought during pregnancy.

In this chapter, I describe the process of identifying variables commonly available in basic EHRs to inform future development of a computable phenotype for pregnancy. Our approach was to analyze a validated pregnant population and compare features to an age-matched control population to identify candidate variables. In future work, these will be incorporated into the development of a multiple logistic regression model that may be validated against a test population and shared with other sites.

Methods

Data Source

As described in chapter 1, all data were extracted from the electronic health record at Nebraska Medicine. These data were transformed to adhere to ONC recommended standard terminologies, fully de-identified, and loaded into an i2b2 infrastructure to allow querying. In the course of this study, additional variables were added to the ETL processes and additional metadata was added to i2b2 to facilitate querying these new data elements. For instance, flowsheet rows regarding fetal measurements were added to the extract. Also, metadata was created to allow for identifying positive and negative pregnancy tests rather than simply the presence or absence of these test.

Candidate Variable Identification

Identifying candidate variables associated with pregnancy was both a stepwise and an iterative process. An initial population with high confidence of pregnancy was identified. This population was used to assess a plethora of variables for inclusion. Clinician review by James R. Campbell, MD (Professor, Internal Medicine Division of General Medicine-Academic) and Teresa Berg, MD (Director - Maternal-Fetal Medicine; Director, Prenatal Diagnostic Center; McGoogan Professor of Obstetrics and Gynecology) pared down the candidate variable list. Based on this analysis, additional variables were extracted from the EHR and the i2b2 metadata was created or refined as necessary to allow querying. Finally, Pearson chi square analysis demonstrated association between identified variables and pregnancy. All elements of this study were approved by the UNMC IRB (#601-17-EP).

Initial Pregnant Population

Within the EHR at Nebraska Medicine, during routine care clinicians may indicate if an encounter is part of an episode. Episodes may be related to surgery, anticoagulation, pregnancy, delivery, or a number of other longitudinal situations. In consultation with domain experts (Campbell and Berg) and in conjunction with a limited chart review, I discovered the flag of pregnancy episode had a high specificity and limited sensitivity for identifying currently pregnant patients. The initial pregnant population was limited to those females ages 15-50 years old with an encounter flagged as part of a pregnancy episode during 2015 or 2016.

Variable Frequency Analysis

For the population identified as described above, all extracted data elements recorded during encounters within the pregnancy episode were identified. These data included vital signs, visit diagnoses, procedures, medications ordered or dispensed,

demographic information, laboratory values, clinical findings, location of visit, and diagnostic related group (DRG) information. The total number and percentage of patients with each fact was identified.

Clinician Review

In consultation with doctors Cambpell and Berg, this survey of variables was grouped and pared down. The initial survey identified each separate diagnostic code and multiple variations of the same laboratory test. The clinicians eliminated non-specific facts (such as common labs for all patients regardless of pregnancy status) as well as grouped similar diagnostic codes at a higher level (i.e. rather than considering Z34.00: “Encounter for supervision of normal first pregnancy, unspecified trimester”, Z34.01: “Encounter for supervision of normal first pregnancy, first trimester”, Z34.02: “Encounter for supervision of normal first pregnancy, second trimester”, and Z34.03: “Encounter for supervision of normal first pregnancy, third trimester” independently, the facts were grouped to become “ENC DX: Supervision of pregnancy - ICD10CM:Z33-34”).

Updating Data Extraction and i2b2 Metadata

This survey of data and clinician review identified two types of gaps in what was previously extracted for obstetric data. First, certain flowsheet rows were missing. Next, the i2b2 metadata allowed for queries to determine if a laboratory test was performed while not allowing querying of the results of the test. To resolve the first concern, the ETL from the research copy of the clinical data research warehouse was updated to include flowsheet information relative to fetal ultrasounds and measurements. In addition, episode data was not originally extracted from the EHR into the CRDW. For these novel data elements, i2b2 metadata was also created to allow for querying.

Next, the analysis demonstrated the metadata previously deployed was insufficient to query laboratory values for various pregnancy tests. For these instances, metadata XML was developed to query the lab values. This necessitated interpreting numeric values as well as free text in some instances to identify each laboratory value as positive, negative, or unknown.

Statistical Analysis

With this enhanced list of candidate variables, contingency tables were created to compare the pregnant cohort to a control population using Pearson Chi-Square tests for significance for each refined variable. The control group consisted of female patients ages 15-50 years old without a pregnancy episode noted during 2015-2016 who had at least one face to face encounter at a Nebraska Medicine affiliated hospital or clinic during this time frame. Patient counts for each population for each variable were identified using i2b2 queries.

Refining the Pregnant Cohort and Control Population for Model Development

Identifying Pregnancy Beginning and End Dates

With candidate variables identified, a verified cohort of all pregnant patients during the date window of interest was required (Figure 5). This needed to include patients with and without pregnancy episodes noted in the EHR. The bounds of a pregnancy were calculated from information extracted from the EHR. The end date for a pregnancy was identified first by the delivery date noted in the obstetric history. If this date was not available, the date of a delivery encounter was used. If this was not identified, the start date of a delivery episode was used to denote the end date of a pregnancy. If none of these were available, the estimated delivery date recorded for the pregnancy episode was used. The start date for a pregnancy was calculated based on

the end date as recorded above. If available, the end date minus the gestational age recorded at delivery was used. When not available, the estimated delivery date – 280 days identified the start of the pregnancy.

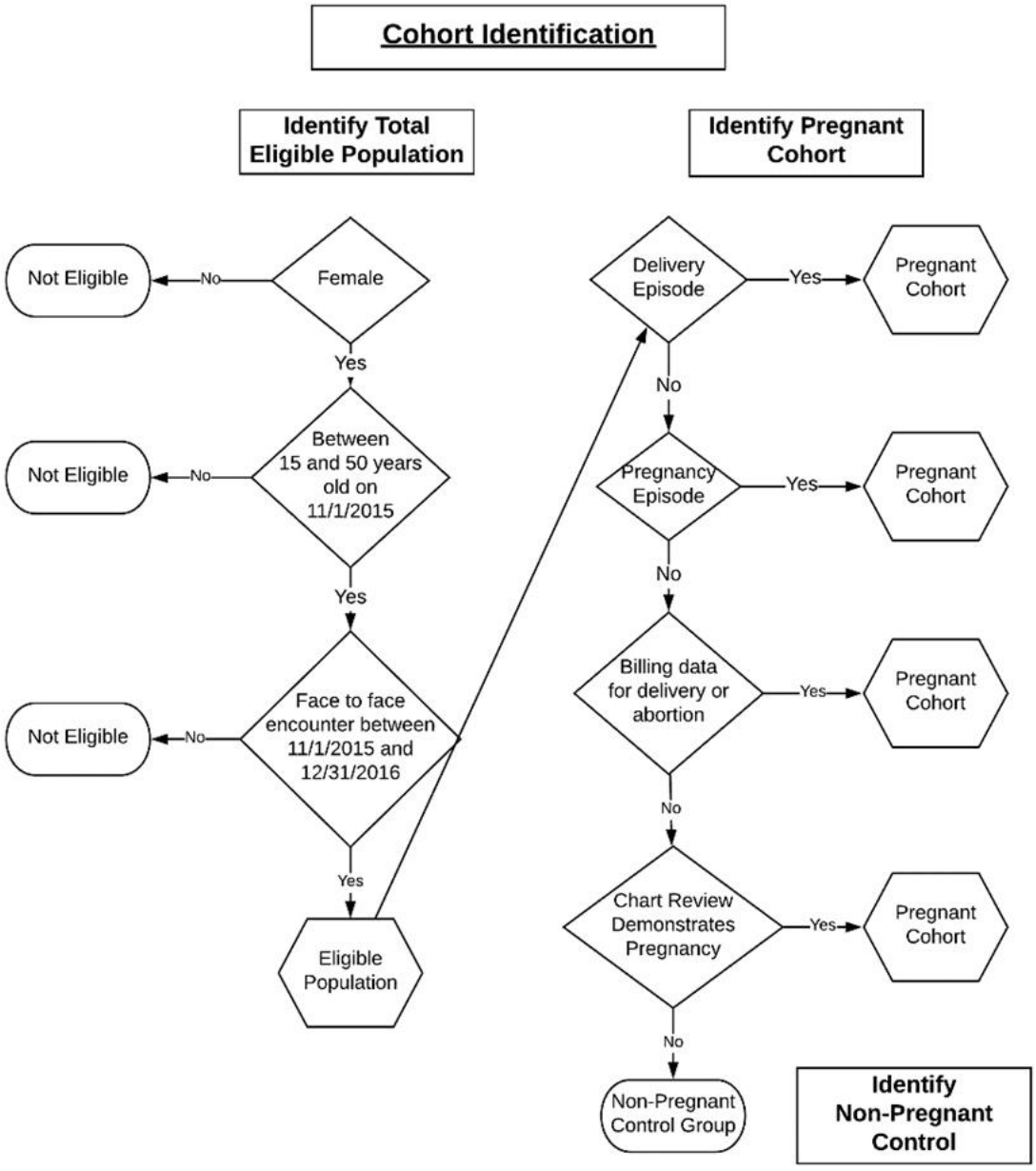

Figure 5. Initial Pregnant Cohort and Control Group Identification Algorithm.

Control Population

Using i2b2, we identified all females ages 15-50 who were seen at Nebraska Medicine for any care during 2015-2016. We excluded patients with a pregnancy episode identified during these years. For both the control and pregnant cohort, patients were required to reside within 50 miles of UNMC. Patients outside of 50 miles often only came to UNMC for specialty care or emergent conditions. During these encounters, pregnancy was noted. However, obstetric care and associated records were maintained elsewhere and not available for inclusion in this analysis. All patients were required to have at least one face to face encounter during the study window.

Results

Variable Identification

Table 6 and Figure 6 provide demographic details about the initial pregnancy episode and control population used to refine the list of potential variables. In both instances, the majority of the population was White and the mean age was approximately 30 years old. Table 7 summarizes the curated list of variables associated with pregnancy. The reported p-values in this table are the results from the Pearson Chi-Square analyses.

Age (years)	Pregnant (n = 5,439)	Control (n = 74,949)
Minimum Age	15	15
Maximum Age	47	50
Median ± SD	28 ± 5.61	31.5 ± 9.00

Race (count (%))	Pregnant (n=5,439)	Control (n=74,949)
White or Caucasian	3,523 (64.8)	55,486 (74.0)
Black or African American	909 (16.7)	7,163 (9.6)
Asian	136 (2.5)	1,975 (2.6)
Other / Unknown	871 (16.0)	10,325 (13.8)

Table 6. Demographics of Pregnant Cohort and Non-pregnant Control Populations.

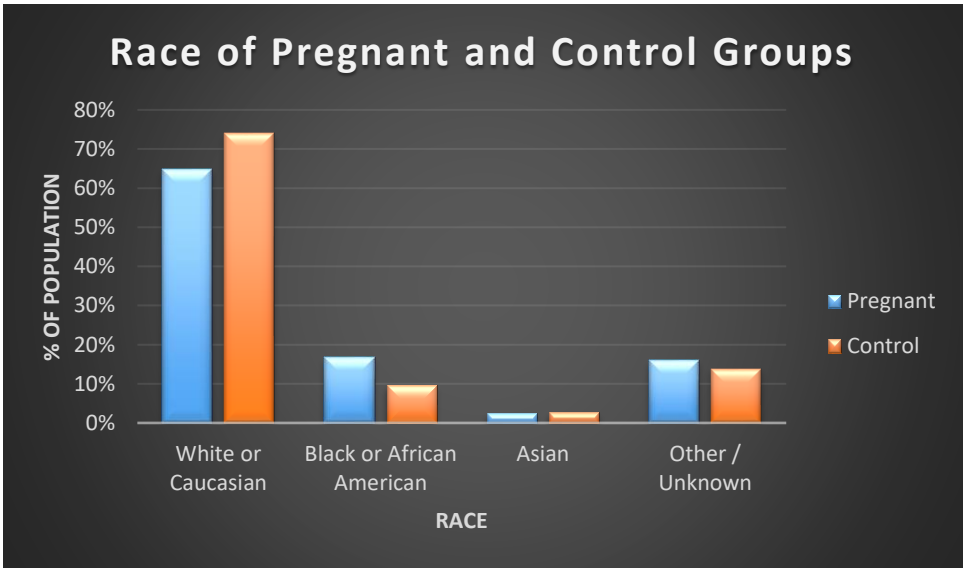

Figure 6. Race of Pregnant Cohort and Non-pregnant Control Population.

Variable	Description	Pregnant (n=5,439)	Control (n=74,949)	p-value
Encounter Diagnoses				
ICD10CM:Z33-34	Supervision of pregnancy	4,408 (81.0)	1,127 (1.5)	<0.001
ICD10CM:O60-O94	Complications of delivery	1,463 (26.9)	374 (0.5)	<0.001
ICD10CM:O009-O29, O98,O99	Disorders during pregnancy	3,915 (72.0)	1,564 (2.1)	<0.001
ICD10CM:O31-O41	Pregnancy complications	1,308 (24.0)	330 (0.4)	<0.001
Procedures				
CPT4:76801-76828	Obstetric ultrasound	4,064 (74.7)	1,285 (1.7)	<0.001
Lab Tests				
LOINC:882-1	ABO and Rh group [Type]	3,513 (64.6)	2,402 (3.2)	<0.001
LOINC:20415-6	B-hCG in serum by immunoassay (Positive)	879 (16.2)	724 (1.0)	<0.001
LOINC:20415-6	B-hCG in serum by immunoassay (Negative)	129 (2.4)	659 (0.9)	<0.001
LOINC:2106-3	hCG in urine (Positive)	178 (3.3)	120 (0.2)	<0.001
LOINC:2106-3	hCG in urine (Negative)	353 (6.5)	6,455 (8.6)	<0.001
LOINC:2118-8	hCG in serum/plasma (Positive)	45 (0.8)	74 (0.1)	<0.001
LOINC:2118-8	hCG in serum/plasma (Negative)	124 (2.3)	2,089 (2.8)	0.03036
Problem List Entry				
SNOMEDCT:173300003	Disorder of pregnancy	3,434 (63.1)	1,293 (1.7)	<0.001
SNOMEDCT:77386006	Patient currently pregnant	3,434 (63.1)	1,293 (1.7)	<0.001
SNOMEDCT:16356006	Multiple pregnancy	110 (2.0)	26 (0.0)	<0.001
OB Clinical Encounter and Measures				
OB Clinical Encounter at UNMC	Visit to obstetric clinic	3,993 (73.4)	9,949 (13.3)	<0.001
LOINC:11881-0	Uterus fundal height tape measure	3,373 (62.0)	337 (0.4)	<0.001
LOINC:55283-6	Fetal heart rate (Positive)	3,552 (65.3)	348 (0.5)	<0.001
LOINC:57088-7	Fetal movement - reported (Positive)	3,194 (58.7)	309 (0.4)	<0.001

Table 7. Comparison of Frequency of Finding in Pregnant vs. Non-pregnant Populations. P-value obtained from Pearson Chi-square test.

Discussion

Conclusions

This survey of the EHR identified a series of variables highly associated with pregnancy. None of these variables taken independently was highly sensitive. Variations in clinical workflow, physician preference, and charting practice within an institution, and likely exacerbated across multiple institutions, may account for differences in data elements being recorded during pregnancy. From initial chart review, some of these variables demonstrated high specificity.

Limitations

A number of considerations must be made when evaluating these data. First, this survey of the EHR was limited to the environment in a single academic medical center with a single EHR. Workflows and technical bias may be introduced as a similar review is conducted at other institutions using disparate, proprietary EHRs. I attempted to mitigate the potential of overfitting to UNMC's environment by identifying a variety of variables from several data domains likely to be recorded in the majority of basic EHRs. I did not simply rely on a single diagnosis or a specific procedure. Rather, data came from procedures, clinical findings, diagnoses, laboratory tests, and obstetric measures. In addition, interoperability and replicability are enhanced as I relied on ONC recommended standard vocabularies to define these variables. With the exception of visiting a UNMC obstetric clinic, all other variables were defined using accepted terminologies.

This review, as well as future creation of a predictive model, was also limited as there is no current and consistent gold standard for identifying a pregnant patient. I used episode data as a specific surrogate for such a standard, however, I noted sensitivity

was lacking. As thousands of patients deliver each year at a Nebraska Medicine hospital, performing a chart review on all patients to be included in the study as well as matched controls becomes a major undertaking. Failure to review all charts may lead to bias as certain individuals are false positives or false negatives as well as having difficulty in identifying meaningful dates for a pregnancy. I attempted to address this challenge by accepting all pregnancy and delivery episodes as being pregnant. As this is a multistep process for a clinician to indicate, the likelihood of false positives is reduced. In addition, I began to refine the population for chart review by identifying patients with high likelihood of pregnancy (i.e. those with multiple, specific findings from the variable list). In this way, future chart review would be limited to those with a lower suspicion of pregnancy as well as matched controls.

Finally, there is the potential for bias as I relied on clinician input to refine the list of variables. Each clinician has a unique background which may affect how she considers characteristics of pregnancy. I limited the potential for bias in two ways. First, I conducted a survey of all recorded facts for patients during the time of interest and examined those with greatest frequency. Any common data elements recorded during pregnancy should be captured in this way. Secondly, I relied on multiple clinicians to review and refine the list of variables. Each could add to or comment on the list to refine it effectively. Bias may be further mitigated as these variables are shared with other domain experts, especially outside this institution, to gain greater perspectives toward general applicability.

Future Research

The methods and results reported in this chapter represent initial stages of a larger effort to develop a computable phenotype for pregnancy status. This initial survey demonstrated variables exist in a basic EHR that are associated with pregnant status.

Immediate next steps will be to continue chart review to identify well-characterized pregnant patients with definitive start and end dates for the pregnancy. Chart reviews will also be required to validate a matched control population for comparison. With these populations identified, each variable will be given an effective time period for which it is an indicator of pregnancy. For instance, a positive urine pregnancy test may indicate a pregnancy six weeks prior to and up to 34 weeks following the date the event was recorded.

With a pregnant cohort and age-matched control identified as well as variables with date ranges associated, a multiple logistic regression approach will be taken to identify a predictive model. The population will be divided into a training and a test group, each with a portion of pregnant and control patients. Random sampling at various time points during the years of interest will be used to fit the model based on the training group. Each patient will have a well-defined period of pregnancy and non-pregnancy. Once refined, this predictive model will be evaluated against the test subset of patients. The effectiveness of the model will be evaluated with a receiver operator characterization curve and its sensitivity and specificity will be noted.

With a well-characterized model, interoperability and extendability will be evaluated. Using the infrastructure and collaboration within the GPC, the variables and model will be shared throughout the distributed research network. This will begin with a survey of sites to determine how many of the 14 variables are currently recorded and extracted from local EHRs. If sufficient data coverage exists, and with IRB approval, a sampling of patients and subsequent chart review will be used to evaluate the applicability and effectiveness of the model.

With this further refinement, the eventual goals are to publish the model in a fashion it may be re-used for clinical research, including pragmatic trials and

observational studies. Further evaluation of the model across larger networks may further refine it. Once widely validated, incorporating the predictive model into the EHR will be planned. Using this model in clinical decision support to identify potentially pregnant patients prior to procedures, medication administration, or imaging will enhance patient safety.

CHAPTER 4 – INCORPORATING A LOCATION-BASED SOCIOECONOMIC INDEX INTO A DE-IDENTIFIED I2B2 CLINICAL DATA WAREHOUSE

Introduction

Clinical research data warehouses are often populated with de-identified patient data extracted from an electronic health record (EHR) [125]. With the continuing advancement and adoption of EHRs, the amount of information available for reuse in clinical research continues to rise [73-76]. However, for complete patient characterization, these data need to be linked to other sources [93]. For instance, while a patient's race, gender, and smoking status are often well-documented in the EHR, other elements of socioeconomic status are often unstructured or absent from the clinical record and unavailable for incorporation into a research data warehouse. These non-clinical elements describing a patient's social, economic, and environmental determinants of health (healthypeople.gov) are a major contributing factor in readmission, morbidity, and mortality [126].

With the paucity of data in the EHR related to socioeconomic status, researchers have relied on insurance type as a proxy for this measure [127]. Data elements related to a patient's neighborhood socioeconomic status (NSES) may reliably be obtained from extra-EHR sources such as American Community Survey (ACS) data from the U.S. census. Neighborhood resources have robust effects on health [128-131], because of their correlation with individual socioeconomic status and as an independent source of influence. The demographic composition of residential areas also has important links to health behaviors and health outcomes [132-135]. Linking measures of the local residential

context to clinical data from the EHR can provide insights into these socioeconomic and demographic correlates of health for researchers.

Using Geographic Information System (GIS) software, a patient's physical address can be linked to a variety of location-based datasets such as the EPA's Air Quality System [136] or the ACS (https://www.census.gov/programs-surveys/acs/). While efforts are being made to integrate these elements directly into the EHR, to date, no EHR has demonstrated widespread integration of such "community vital signs" [137]. Implementing this linkage for clinical research with an EHR-agnostic approach introduces additional challenges related to patient privacy and data standardization. The Health Information Portability and Accountability Act (HIPAA) privacy rule, the Health Information Technology for Economic and Clinical Health Act of 2009, and the Federal Policy for the Protection of Human Subjects are designed to safeguard protected health information (PHI), including street address and geocodes [138]. These safeguards may prohibit researchers from sharing the information required for geocoding with an academic or business partner third party, hampering the ability to link clinical and NSES data within an institutional review board (IRB) approved process. Additionally, data that would identify a patient's location with too much granularity may not be displayed in a de-identified data warehouse. For instance, HIPAA requires zip codes be obfuscated if the population is below 20,000 for that area. Many details are available from the ACS for significantly smaller populations, requiring obfuscation before being made available in a de-identified system.

One approach to maximize sample population while maintaining patient privacy is to participate in a distributed research network (DRN). The National Patient-Centered Clinical Research Network (PCORnet) and its participating Clinical Data Research Networks (CDRNs) illustrate how patient data may be stored locally and federated

queries may be shared across the network [2, 97, 100] . However, interoperability of research queries across a CDRN is challenging as variability may be introduced into a collaborative study if geocoding is performed independently at each site with disparate methods [139-141] . This variability may affect analysis and conclusions in health care studies [142] . As clinical and socioeconomic data are linked, successful collaboration and data analysis is dependent on a means of querying these data from each site for a variety of studies.

In this chapter, we provide a model for combining socioeconomic and clinical data while maintaining patient privacy and allowing rapid querying in a de-identified data warehouse. We describe an algorithm for extracting socioeconomic status data from the American Community Survey (ACS), geocoding patient data without involving a third party, and combining these data within an Informatics for Integrating Biology & the Bedside (i2b2 - www.i2b2.org) framework for interrogation. Due to the volume and variety of data within the ACS, we extracted only elements to calculate a validated socioeconomic index [126] . We demonstrate this data extraction and incorporation into the research infrastructure using an example of evaluating the impact of NSES on emergency department (ED) utilization. The approach we describe may be fully deployed at other sites and will allow for collaborative research and federated queries while keeping PHI secure [143]

Methods

Clinical Data

Patient data were extracted from a research copy of the EHR data warehouse at the University of Nebraska Medical Center (UNMC). This system contains data originating from multiple hospitals and clinics in urban, suburban, and rural Nebraska.

Clinical and demographic data are extracted, standardized based on Office of the National Coordinator (ONC) recommended vocabularies, and transformed for use within an Informatics for Integrating Biology & the Bedside (i2b2) clinical data warehouse [118, 119]. Data are transformed and staged on an identified server and then de-identified and made accessible to researchers via i2b2 in a fully de-identified database on a separate server (Figure 7). This data extraction and use in a de-identified data warehouse was approved by the institutional review board (IRB) at UNMC (IRB #132-14-EP).

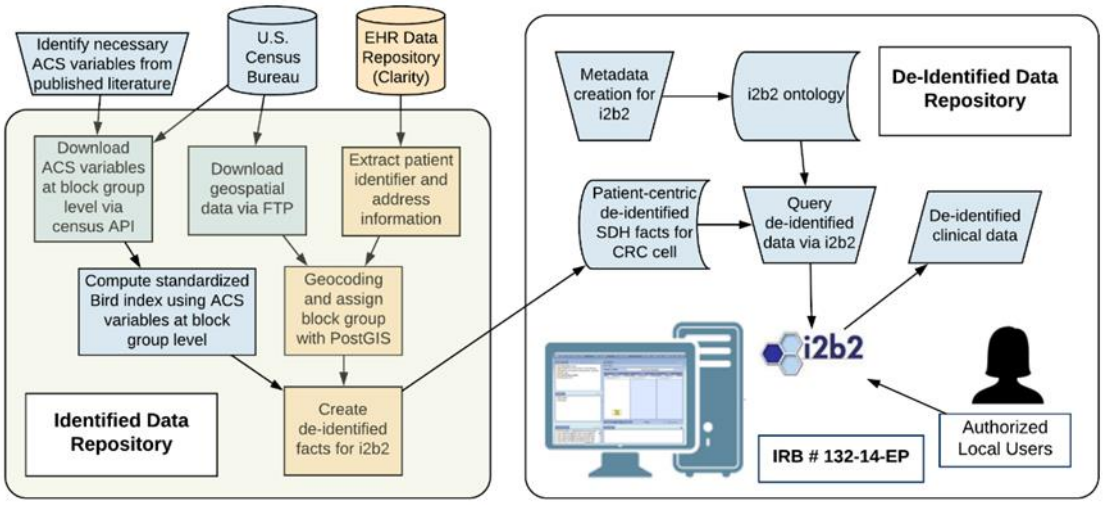

Figure 7. Integrating Census Data into a De-identified Data Warehouse and Querying with i2b2

Geocoding Process

Current and historic patient address information was extracted from the EHR and stored on a secure server. TIGER/Line Shapefiles and other location-based files needed for geocoding were obtained via FTP from the United States Census Bureau and loaded onto the server alongside patient address data. These files were for year 2017 and for the states of Nebraska and Iowa. Using these data, PostGIS version 2.4 [144] geocoding software running on PostgreSQL version 10.1 was used to identify the longitude and latitude for each patient address. Subsequently, the U.S. census block group for each successfully geocoded address was determined via PostGIS. For this study, we geocoded only patient addresses for Nebraska and Iowa. In the extraction, we eliminated P.O. Box addresses and those addresses with null or invalid street addresses (Figure 8). Invalid street addresses were defined as those consisting of all alpha or all numeric characters, a single character, or containing variants of the words unknown or invalid.

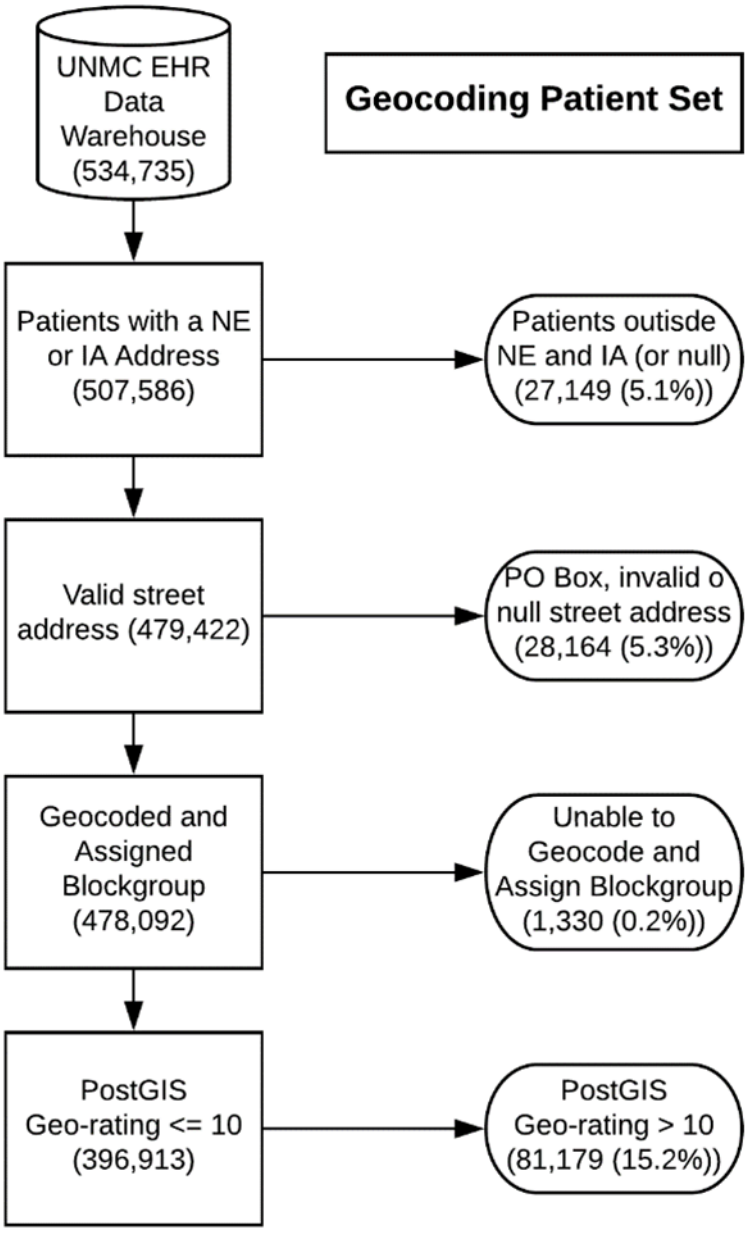

Figure 8. Identifying patients with a well-geocoded address

We compared demographic data for the geo-coded population relative to those patients we excluded from analysis. Rural versus urban location was based on United States Department of Agriculture (USDA) Rural Urban Commuting Area (RUCA) codes mapped to zip codes (https://ruralhealth.und.edu/ruca , https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/documentation/). Financial class for the patient was defined as the primary insurance category listed on the patient's account in the EHR. Age was calculated from the date of the data extract (December 2017).

Census Variable Extraction and Socioeconomic Index Calculation

Based on the index and variables described by Bird et al, a set of equivalent variables which could be computed from the U.S. Census Bureau's annual American Community Survey (ACS) were identified [126]. Table 8 identifies the Bird variables from the 2000 decennial census and the field and computation employed from the 2011-2015 five-year estimates from the ACS. Using the U.S. Census Bureau API, ACS estimates for each variable for each block group in the U.S. were extracted and stored locally. These raw estimates were transformed and normalized for all Nebraska and Iowa block groups to have a mean of 0 and a standard deviation of 1 as described by Bird [126]. The Bird NSES is computed as the sum of the standardized values for each of the six variables, where the standardized values are multiplied by -1 for variables where a higher positive value indicates lower socioeconomic status. This method results in higher Bird index values corresponding to a higher socioeconomic status. If any of the six variables were unavailable from the ACS for a specific block group, the Bird index was not computed for that block group. Standardized values for the six variables as well as the Bird index were stored for all block groups within Nebraska and Iowa.

Description	ACS Variable	Computation
Percent of adults older than 25 with less than a high school education	B15003	$\frac{\sum_{i=2}^{16} HD01\_VD_i}{HD01\_VD01}$
Percent male unemployment	B23022	$\frac{HD01\_VD26}{HD01\_VD02}$
Percent of households with income below the poverty line	B17017	$\frac{HD01\_VD02}{HD01\_VD01}$
Percent of households receiving public assistance	B19057	$\frac{HD01\_VD02}{HD01\_VD01}$
Percent of female-headed households with children	B23007	$\frac{HD01\_VD26}{HD01\_VD01}$
Median household income	B19013	$HD01\_VD01$

* Data extracted from the 5 year estimates for 2011-2015.

Table 8. Description of Variables to compute neighborhood socioeconomic status

Identifying Patient NSES and i2b2 Fact Creation

Patient addresses from Nebraska and Iowa were linked to census block groups. Patient identifiers were linked to the Bird index and standardized ACS variables associated with their block group. Data were de-identified and loaded into the database on the de-identified server used by i2b2. De-identification included using a randomly generated patient number, shifting all dates for each patient randomly by -1 to -365 days, and excluding any HIPAA identifiers. For each block group, we ensured the level of granularity of the reported NSES index was only sufficient to match at least seven block groups. This ensures the patient remains anonymized in a population of at least 20,000 people. For each patient, seven records were inserted into the database: the Bird NSES index and the six standardized variables necessary for its computation.

Metadata Creation and i2b2 Querying

The demographics portion of the i2b2 ontology cell was updated to support interrogation of ACS based facts (Figure 9). A folder for neighborhood socioeconomic status was added with subfolders for the Bird NSES index and the six, standardized variables of interest. ACS variables were identified with the field label as well as a brief description to ensure they are both standardized and human readable. Metadata XML was included to allow users to specify ranges or values of interest for each of the variables.

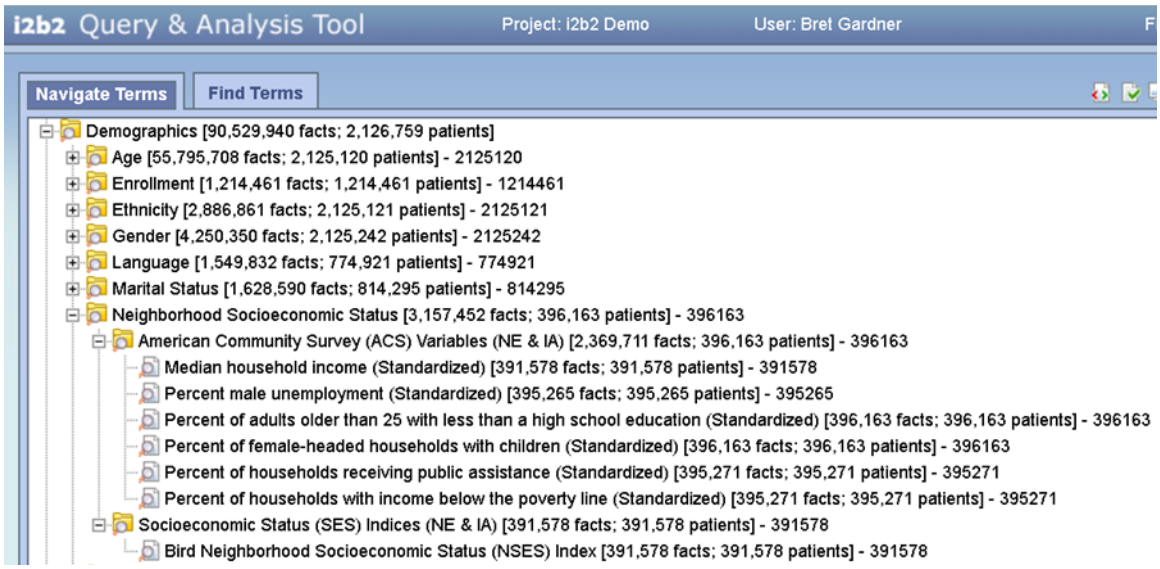

Figure 9. Integration of ACS Metadata into the Demographics Hierarchy of an i2b2 Client

Example Use Case: Emergency Department Utilization

All patients with a computed Bird socioeconomic status index who were seen in any Nebraska Medicine affiliated hospital or clinic between January 2013 and December 2017 were identified. An odds ratio was computed comparing the number of patients with zero emergency department (ED) visits to those patients with one or more ED visits during 2013 to 2017. These patients were stratified on the basis of above average or below average Bird index. ED visits included encounters with any resulting discharge disposition, including hospital admission or expiration. In addition, for patients with an ED encounter and two or more total face to face encounters within the target date range, an ED frequency metric was computed. For each patient, the total number of ED encounters was divided by the time between the earliest and latest face to face encounter of any type. A one-tailed student's t-test with alpha of 0.05 was used to compare this metric between the high and low NSES populations (R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.). Results

Results

Geocoding

We geocoded only patients with a Nebraska or Iowa address, representing the majority (507,586 / 534,735 (94.9%)) of Nebraska Medicine patients. Patients were excluded from analysis if: 1. The patient address failed to geocode or have a block group assigned (1,330), 2. The patient street address was unknown or invalid (28,893), or, 3. The geo-rating assigned by PostGIS was greater than 10 indicating a low confidence in the geo-code assignment (81,179). The final geo-coded population consisted of 396,913

patients who have had an encounter at Nebraska Medicine, a well-geo-coded address, and a block group assigned (Figure 8).

Figure 10 illustrates the comparison of the included versus the excluded population for analysis. The percentage of the excluded population living in a rural zip code was 23.7% compared to only 9.8% of the included population. The racial composition of the included and excluded populations were very similar with each demonstrating a majority white (78.9% and 80.8%, included and excluded populations, respectively) with a lower percentage of black (9.5% and 7.6%) and other races (11.7% and 11.6%) in the both populations. The populations demonstrated little difference in gender proportion (54.7% and 52.7% female, inclusion and exclusion population, respectively). The included population had a higher percentage of private / commercial insurance (48.09% vs. 45.44%) while having a slightly lower Medicare percentage (16.5% vs. 20.1%). The included population had a lower age relative to the excluded population (age years (SD)) (42.1 (23.6) vs. 44.6 (24.0)).

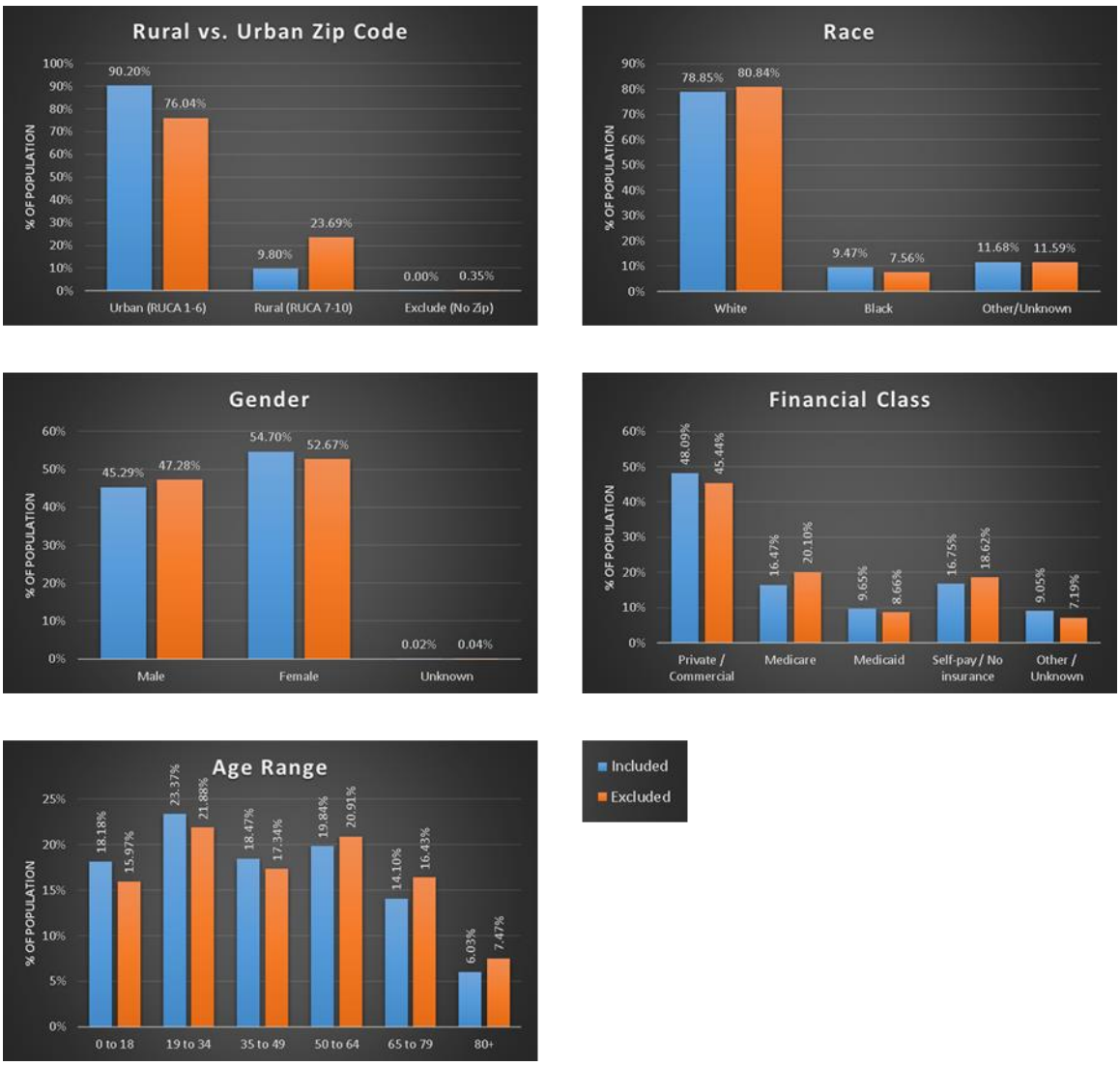

Figure 10. Comparison of Geo-coded and Excluded Patient Population

Neighborhood Socioeconomic Status Variables

For the six variables we computed from the ACS data, values were present for at least 98% of block groups in Nebraska and Iowa. We were able to compute a Bird index for 4,208 of 4,263 (98.7%) of all block groups in these states.

ED Utilization

Using the i2b2 data warehouse, 360,947 patients were identified as being seen in any Nebraska Medicine hospital or clinic between January 2013 and December 2017 who also had an assigned Bird index. Of these 214,325 (59.38%) had above average index values and 146,622 (40.62%) had below average values. For patients with a below average index value for their neighborhood, 60,309 (41.13%) had at least one visit to an emergency department during the study period for any reason. During the same period, 59,550 (27.78%) of patients with an above average index had an emergency department encounter (Figure 11). Patients living in an area with a below average index have 1.82 times the odds of visiting the emergency department compared to patients with living in neighborhoods with a higher index (95% CI 1.79-1.84).

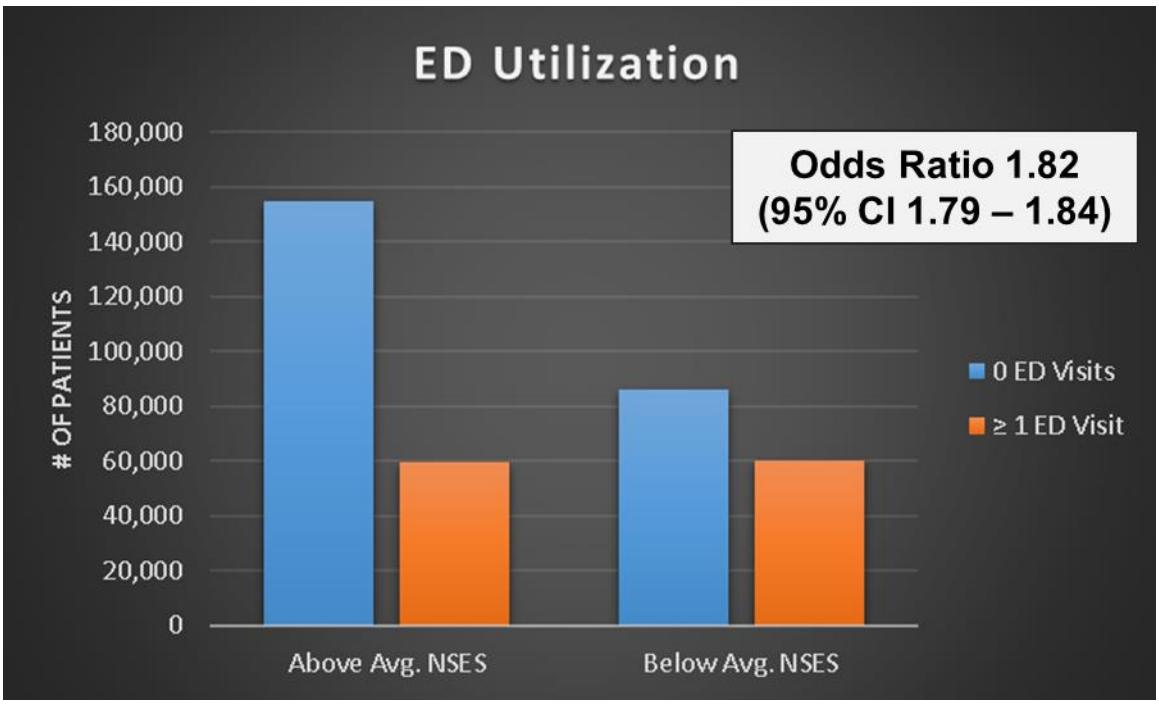

Figure 11. ED Utilization Stratified by NSES

For patients with at least two total face-to-face encounters and at least one emergency department visit between 2013 and 2017, there were 39,697 patients residing in areas with high NSES and 39,535 patients in areas with below average NSES. Figure 12 displays the distribution of ED utilization rates for each of these populations. The average number of visits to the ED per year were 2.11 for patients with an above average NSES and 2.40 for patients with below average NSES. A student's t test demonstrated that patients with below average NSES had a mean emergency department visit rate of at least 0.2 visits per year higher than those with above average NSES (alpha=0.01, $p < 0.001$).

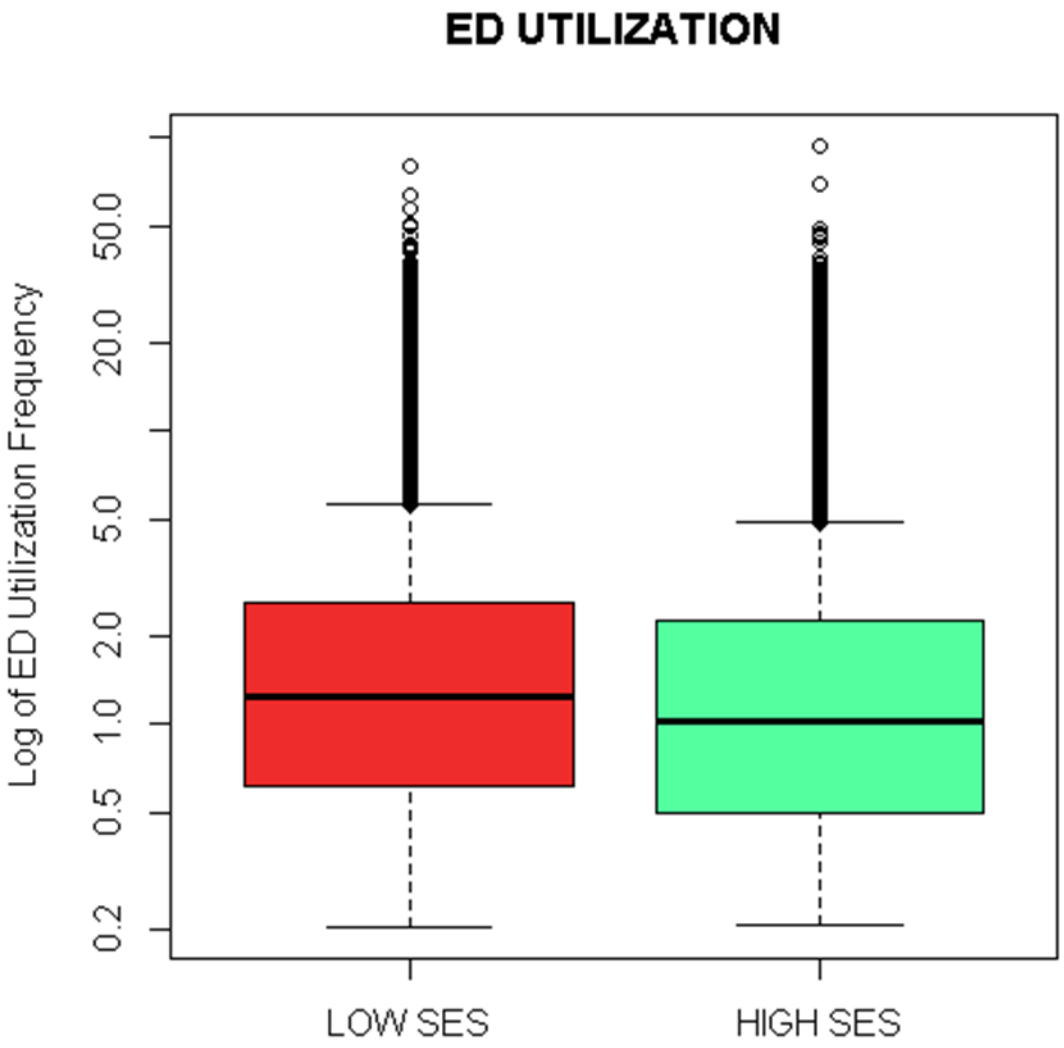

Figure 12. ED Utilization Rate Stratified by NSES

Discussion

Conclusions

With evidence of the impact of environmental factors on health, facilitating comparative effectiveness clinical research incorporating social determinants of health is paramount [145]. Advances in geoinformatics make it possible to link patient locationto data provided by the U.S. Census Bureau. We demonstrated an approach to link social determinants of health data from the American Community Survey to clinical data in a de-identified data warehouse. All elements of this approach may be completed at individual sites, avoiding the need to send PHI to a third party. When sites utilize the same geocoding and linkage process, collaborations are possible without introducing unnecessary variability between sites. Institutions who implement this approach using i2b2 may share federated queries across networks such as PCORnet and the GPC, to increase the patient sample size for analysis [146]. Facilitating this research will inform efforts to incorporate location-based census data directly into the EHR and future clinical decision support (CDS) at the point of care.

This study is an example using only a single socioeconomic index. While many indexes have been published to estimate socioeconomic status, we selected Bird's model as it is well-validated and fully reproducible using data elements from the ACS [147-151]. Future work includes incorporating other well-validated models based on extant ACS data using the process described in this manuscript. These may readily be incorporated into the database within the ontology cell of i2b2. We demonstrated the efficacy of querying these data with a use case based on evidence from prior studies. We noted both a higher proportion and more frequent per patient utilization of the ED for patients

living within areas of lower NSES relative to patients from areas of above average NSES.

Limitations

This study is limited by only including patients from Nebraska and Iowa. Additionally, the clinical data utilized does not encompass all hospitals and clinics patients may visit. Integrating health information exchange data would enhance the clinical picture and facilitate studies investigating readmission.

This study is also limited by the quality of data available in the EHR. For instance, 28,893 / 508,315 (5.68%) of all addresses were unable to be geocoded as they were recorded as some variant of unknown, were null, or were P.O. Boxes rather than a physical address of a residence. In addition, there is varying quality of confidence in the results returned by the geocoding software with 81,179 / 508,315 (15.97%) having a geo-rating with low confidence (PostGIS geo-rating > 10). As addresses are non-uniform and may contain errors, some may not geocode accurately. By excluding patients and potentially mismapping a small portion of the patient population, the potential for bias is introduced. While race and gender showed no significant difference between the included and excluded populations, as is evidenced in other studies, rural locations had a lower percentage of successful geocoding [152]. Differences in these populations were also seen for age (included population slightly younger) and financial class (Medicare patients more likely to be excluded). Recognizing these population differences is essential as they may impact analyses when future studies rely on these data [153]. While the geocoding for this study did not reach 100% completeness or 100% accuracy, results were comparable with other first-pass geocoding efforts [152, 154-159]. A refinement of the geocoding process may reduce this bias and increase confidence of results relying on the generated data. Perfecting geocoding is beyond the scope of the current

demonstration. An enhanced geocoding process could readily be integrated into this model for incorporating census data into a de-identified data warehouse.

Future Research

To demonstrate reproducibility and extend the results demonstrated in this paper, collaborations to implement this data integration approach will occur in existing distributed research networks (Greater Plains Collaborative (GPC) and The National Patient-Centered Clinical Research Network (PCORnet)). Within the GPC, sites have implemented de-identified data warehouses on an i2b2 platform. Across PCORnet, each participant transforms clinical data to adhere to the common data model (CDM). While the i2b2 querying approach will not be able to be demonstrated throughout PCORnet, the CDM and established federated query protocols will allow collaboration.

Future efforts will also address incorporating additional socioeconomic indices and ACS variables into the de-identified clinical data warehouse [148, 150, 160] . As part of this effort, a standard approach to identifying these indices and component variables within an i2b2 ontology will be proposed for interoperability. Integration of additional indices will allow both the comparison of the efficacy of indices in a variety of contexts as well as the application of validated indices within many disease phenotypes.

Acknowledgements

The project described was supported in part by the National Institute of General Medical Sciences, 1U54GM115458-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

This work is partially supported through a Patient-Centered Outcomes Research Institute (PCORI) Program Award (#CDRN-1306-04631).

DISCUSSION

Summary

Overview

In this dissertation I have demonstrated an approach to assessing the utility of a clinical research data warehouse for effectively and efficiently replicating clinical research. With the gaps identified in this assessment, I also performed preliminary work toward developing a computable and shareable phenotype for pregnancy status based on EHR data. Finally, I developed and deployed a methodology for linking EHR data to socioeconomic data contained within the U.S. Census within an i2b2 infrastructure. While replication of all studies remains a future goal, the results described in this dissertation represent progress toward this end and provide areas of consideration as institutions establish clinical data research warehouses or participate in distributed research networks.

Assessing Risk Models and Infrastructure for Replication

First, using two risk models based largely on data from basic EHRs, I assessed their replicability with the infrastructure established at UNMC. While I surveyed a variety of risk models, those included in this dissertation clearly demonstrated significant points in the replication process. First, these models were clinically meaningful and based on EHR data. While not always the case, these two models each provided meaningful results for a clinically significant concern (readmission in heart failure patients and risk of in-hospital mortality). If the initial results are not clinically significant, attempting to replicate the model may be superfluous. This may be due to being statistically significant but not clinically meaningful (i.e. less than a one-point systolic blood pressure difference)

which may be detectable with very large sample sizes, or, due to being relevant to only a very specific subset of the population that may not be represented at all sites (i.e. a very rare disease or an effect linked to geographic region). Next, each model was largely based on data likely to be found in a basic EHR. Other models may be based on data from surveys, disease registries, or public records. While replication of these models may be possible, this would require incorporation of novel data sets for each model rather than clinical data that may be re-used for a plethora of models. The generalizability and replicability of risk models relying on extra-EHR data is hampered by the resource cost of instantiating these data.

Next, beyond the clinical significance and the reliance on EHR data, to assess the utility of UNMC's CRDW to replicate risk models, I surveyed the specific data elements the risk model was built on. Each of the risk models in this dissertation were designed to largely use data collected on the majority of patients within the first 24 hours of hospital admission. UNMC's CRDW demonstrated excellent coverage of demographics data and a large portion of encounters had laboratory data as well. One element limiting the replicability of the heart failure risk model was the inclusion of additional data elements, such as socioeconomic status based on census data, number of address changes in recent history, and the number of missed visits. While visit and address information are recorded in the EHR, these may not be commonly extracted into a clinical research data warehouse. Developing ETLs for specific data elements requires an investment of time and resources. Prioritizing these extracts to the data elements used most frequently is essential. In addition, extra-EHR data, such as that found in the U.S. Census may be linked to EHR data, however, as demonstrated later in the dissertation, significant resource investment is required. One contribution of these efforts toward replication was the prioritization of future data extracts into UNMC's data

warehouse. Assessing specific data elements is essential to evaluating a site's ability to replicate a study.

An essential element of this data survey is the definition of each data element. For a significant amount of clinical data, the ONC has recommended standard vocabularies for use in recording and describing data. If employed and well-documented, these standards facilitate interoperability of healthcare data as well as enhance replicability of published research. For the two risk models described in this work, neither fully described all clinical elements with these recommended vocabularies. Amarisingham et al had a well-defined phenotype for the heart failure patients based on ICD-9 CM codes. However, in both studies, only the laboratory name was given rather than a series of LOINC codes. This necessitated the replication effort to include a lookup of a significant number of laboratory codes and may represent a departure from the original studies. In neither case was the phenotype or inclusion criteria published in a machine parseable form. My work involved manually parsing the descriptive text and interpreting definitions for each data element. To enhance replicability to make this effort more efficient, authors may publish a human readable description in conjunction with a fully machine-readable document using standard terminologies.

In addition to ensuring data elements are present in the data warehouse, ensuring data volume is sufficient for the study is critical. I assessed data volume for each model by developing and executing a series of queries in i2b2. This provided simple patient counts for each variable of interest. Some data may be extracted but exist at a very low rate. For instance, while all laboratory elements for the heart failure model were extracted from the EHR, no patients in this cohort had a pro-BNP recorded. For a rare disease, it is possible an insufficient number of patients may exist at a single institution necessitating collaboration across a network. This was not a concern for the

two models described in this dissertation as the i2b2 queries demonstrated a large patient population for the common cohorts of heart failure and hospital admission. Ensuring both data variety and data volume are important elements in ensuring a site has the capability of replicating published research.

Along with assessing the clinical relevance and potential generalizability of a risk model and surveying the specific data elements required in the model, an evaluation of the statistical methods the authors used is critical when determining if a model may be replicated. While Amarasingham's risk model for readmission in heart failure patients may have demonstrated significance, neither the publication nor attempts to reach the author were sufficient to elicit a replicable set of methods. While data limitations precluded replication of this model, even if all data were present, the absence of clear methods prevents replication of the work. At best, I could demonstrate correlation between the same set of variables and readmission rate. It would be guesswork to determine if the variable had relative contributions as described by the authors. Whereas the methods for the first model remained somewhat a black box, Tabak's mortality risk model had well-described analyses and model development. Prior to investing significant resources toward extracting additional data elements or linking to external data sources, it is essential to ensure the analyses are well-described and well-designed.

To summarize, any attempt toward replicating a published risk model with clinical data requires assessment both of the published work and of the infrastructure at the new site. In this dissertation I demonstrated these processes with two separate models. In each instance, I ensured the model had clinically significant results, was applicable to the population at this site, and published sufficient methods to allow me to perform similar analyses and model development. In addition, I surveyed data available within the clinical research data warehouse to ensure volume and variety sufficient to model

the methods in the original work. Neither paper published fully defined criteria using standard terminologies. One model failed to publish a complete description of the methods. This model also employed extra-EHR data not available within UNMC's data warehouse at the time of the replication attempt. My overall assessment indicated one of the two studies may be replicable at UNMC. The other study may have valid results, however, due to infrastructure limitations and publication omissions, replication is not possible.

Replication Attempt of Risk Model

With the model assessment in place, an attempt to replicate Tabak's mortality model was possible. In order to replicate this study, definitions of data elements needed to be interpreted, local data needed to be queried, and analysis of the data as described by the original study was required.

As described above and documented in Appendix C, I translated the narrative description of laboratory values and inclusion criteria into LOINC codes to interrogate the clinical data at UNMC. Multiple LOINC codes were possible for the majority of laboratory tests described. Through i2b2 queries I answered which codes were used in the EHR and populated in the data warehouse. In consultation with James R. Campbell, MD, a clinician and the institution's terminologist, I ensured the codes I selected were comprehensive and accurate. Once defined, these codes are shareable to any other site desiring to replicate this or a similar study.

With standard codes in place, clinical data needed to be queried. While feasibility counts for the study were obtained using i2b2 queries, data extraction required SQL scripts to be written. Scripts were designed to identify the inpatient encounters of interest, extract raw values for each of the necessary laboratory values, and calculate a total score for each patient encounter. These scripts rely on standard terminologies and

the i2b2 star schema data structure. They may readily be shared and adapted to other sties who have implemented i2b2. In this way, this work may be re-used to facilitate future replication.

Finally, with standard definitions in place and re-usable SQL scripts developed to extract the data, the methods described in the study were replicated. Using R, a precision recall as well as a receiver operator characteristic curve were created to assess the predictive value of the total calculated score. This demonstrated similar predictive power to what was described by the original study. In addition, multiple logistic regression analysis was used to develop a new risk model using the same variables. This model demonstrated similar predictive power to the original model.

Through the assessment described above, I demonstrated that a replication attempt using the Tabak mortality risk model was possible. With the data extraction and statistical analysis performed in R, I demonstrated replication of the study. The published results were validated by this work and extended to a novel population. Future replication at other sites is possible as standard definitions are now in place and data extraction and analysis code is available for re-use.

Creating a Computable Phenotype for Pregnancy

In this dissertation I described initial steps toward creating a computable phenotype for pregnancy status based on data commonly recorded within an EHR in the course of typical care. Based on a number of queries from participation in a distributed research network which had variable or poor definitions of pregnancy, the need for a consistent and interoperable phenotype was recognized. While fully-defining, validating, and sharing such a phenotype will take a greater investment of resources and input from a variety of experts, this dissertation demonstrates a series of variables for incorporation into a fully-defined phenotype. Through frequency analysis, expert input, and statistical

analysis, a number of data elements were seen to correlate with being pregnant. This dissertation provides definitions for these data elements using standard terminologies (LOINC, SNOMED-CT, CPT, and ICD*) to enable interoperability of the definition. This work also demonstrates how these variables may be extracted from the EHR as well as how to identify temporal bounds for pregnancy from the EHR.

Incorporating Location-Based Data into the i2b2 Infrastructure

As with the necessity to develop a computable phenotype for pregnancy, the need for a means of incorporating queryable location-based data into the clinical research data warehouse stemmed from working with other studies. As described above, Amarasingham's risk model for readmission in heart failure patients requires socioeconomic variables available within the U.S. Census. While many studies rely on these data, the novel contribution of the results presented herein is a demonstration of a replicable means of incorporating these data into an i2b2 infrastructure in a de-identified data warehouse.

Geocoding has greatly advanced in recent years and there are countless approaches to translating addresses into longitude and latitude and later linking this to census data. One important element of the approach taken in this dissertation is the geocoding was all done in house. Patient data remained on a secure server, never requiring a third party or any data transfer which may increase the risk of exposing PHI. While enhancements to geocoding are beyond the scope of this dissertation, demonstrating an approach to securely geocode patient data in an IRB approved manner was critical. Of note, alternative geocoding processes may readily be implemented into the pipeline described in this dissertation.

Beyond demonstrating a secure approach to geocoding patient addresses, this dissertation demonstrates a method for maintaining HIPAA compliance while exposing

patient neighborhood socioeconomic status. HIPAA geographic requirements mandate de-identification ensures that a patient's geographic location may not be identified to a catchment area with fewer than 20,000 people. We de-identified the data in a number of ways. First, rather than publishing raw data from the ACS, standardized data and summary indices were displayed in the i2b2 environment for researchers. As we publish our methods and as other researchers could potentially work back to which blockgroup a standardized variable originated in, we also truncated these published variables to ensure a sufficient number of blockgroups are identified with any given published result. In this way, only a group of blockgroups could ever be identified for a single patient, ensuring at least 20,000 people were located in this geographic region.

To demonstrate the validity of this approach, we replicated a well-published result of ED utilization. Using i2b2 queries for feasibility queries and SQL scripts for data extraction, we demonstrated increased ED utilization based on residence in an area with below average socioeconomic status. While this result is not novel, our work serves to validate the result on a novel population. Furthermore, our results demonstrate the efficacy of our approach to incorporating socioeconomic variables in a fully-deidentified i2b2 data warehouse. Our approach is both efficacious and replicable.

Assessment of Hypothesis

The methods and results presented in this dissertation formed to evaluate the hypothesis that the data and infrastructure of UNMC's clinical research data warehouse were sufficient to allow for replication of risk models and other clinical studies. As was demonstrated early on, this hypothesis was rejected and required modifications. Early work demonstrated inability to replicate studies was due to two major categories. First, insufficiency of extracted data types in the CRDW, and, second, insufficiency of

published methods and analyses to allow any site to replicate the study. The first category led to the work toward developing a shareable, computable phenotype for pregnancy and the process of incorporating socioeconomic variables from the census into the data warehouse.

From the results described in this dissertation, a refined hypothesis that may undergo continued testing is the CRDW infrastructure is sufficient to allow replication of studies publishing clear methods and analyses. Some studies may require investment of resources to extract additional clinical variables or link to extra-EHR data sources. The assessment pattern described herein allows rapid identification of the replicability of a study and identification of any missing variables requiring ETL development. In this way, if a study may be replicable, additional data elements can be prioritized in the extraction workflow.

Generalizability of the Results

The results documented in this dissertation are not without limitations. This work was done in a clinical data research warehouse based on EHR data with some transformations. These data were standardized to ONC recommended vocabularies, either within the EHR or after extraction. In addition, these data were de-identified during extraction, precluding incorporation of some identifying variables per HIPAA guidelines. Finally, these data were stored in a form to foster interoperability via i2b2 as well as the PCORnet CDM. The standardization made the data survey for required elements much simpler. The methods described herein to assess the utility of a CRDW for replication of a study may be more challenging to employ without standard vocabularies in place.

The results of replicating the risk model for in-hospital mortality are expected to be highly generalizable. With the methods available, the specific variables well-defined,

and the model validated on two independent hospital environments, other sites with a typical hospital census should find high utility in the model. Generalizability could be further validated by evaluating the mode's performance across the GPC or all of PCORnet's DRNs.

The methods and results reported regarding incorporating location-based data into a de-identified clinical data warehouse should also be generalizable. The ED utilization result was validated in a novel population with novel calculation of socioeconomic status. As with the risk models reported on, further validation and generalizability may be achieved if this hypothesis were tested across a DRN rather than at a single site. The methods to incorporate and use these data were designed to be portable and reproducible. Ideally, other sites could implement this workflow and have queryable variables in a de-identified i2b2 data warehouse.

Future Work

The results and methods presented in this dissertation facilitate both immediate next steps for research and long-term expansions of these concepts. Future research is possible focusing on three distinct areas. First, further validating the risk models described herein as well as additional risk models. Next, continuing the development of a computable phenotype for pregnancy status and sharing this phenotype. Finally, incorporating and utilizing additional location-based variables from the ACS into the de-identified i2b2 data warehouse.

Future Research Replicating Risk Models and Observational Studies

In-hospital Mortality Model

First, the risk model for in-hospital mortality may be further validated and used as a test case for replication research across a distributed research network. I demonstrated a means of assessing the replicability of a study at a single site. I also validated this previously published risk model. With similar methods, an assessment of the replicability of this model across the GPC is possible. Shared i2b2 queries or SQL may be minimally modified to survey other sites to ensure data volume and variety are queryable relative to this risk model. Independent models could be constructed at each site, or, with IRB approval, de-identified encounter level data could be analyzed at a central site to test the validity of the model on a larger population covering much of the Midwest. While greater adaptation would be required, SAS scripts or SQL could be made to query the PCORnet CDM. If a data survey indicated the necessary information is currently populated in the CDM, any participating site across the country could execute the queries and the model could be replicated at dozens of sites.

Potential Problem of Heterogeneity

Attempting to replicate risk models and observational studies across diverse distributed research networks offers advantages of large sample size and a more diverse study cohort. However, as described earlier, there is evidence this may introduce bias due to the heterogeneity of the databases and environments being combined. One area of necessary future research beyond the scope of this dissertation is to determine if bias is introduced from pooling diverse data sets. If so, does this significantly affect the outcomes of federated queries? Will pooling the data across many sites mask individual variations in results, or, does conforming to the common data model allow the same results to be observed with greater statistical significance?

Higgins et al. proposed a means to increase the efficacy of meta-analyses, taking into account the potential heterogeneity of studies being included [161] . While this is

certainly valuable and will add credence to the conclusions of meta-analyses, further efforts to provide a means to validate the results of collaborative studies from various medical centers is needed to ensure individual variation is not simply being averaged and true results lost. Madigan et al performed such a study in the OMOP DRN. Attempting to replicate that analysis across the GPC or across PCORnet will provide valuable data about potential bias. The results of such an analysis will have significant impact on future use of data from large networks.

Additional Study Replication

The work presented in this dissertation represented a survey of a series of risk models and efforts to replicate only two of them. Countless risk models and observational studies have been published while very few have been replicated or validated. Applying the principles outlined in this dissertation will facilitate a researcher's effort to determine if a study is replicable and what resource investment is required to allow replication at any site or across a network. This process may be followed for specific populations, such as those with a given disease, or for a variety of risk models ranging from mortality to other outcomes. Such replication studies will direct resource utilization in incorporating further data into research data warehouses as well as identify valid risk models that may be incorporated into an operational EHR for CDS.

Future Research Developing a Computable Phenotype for Pregnancy Status

Temporal Variable Definitions

The results described in this dissertation represent the foundation for future research in developing a computable phenotype for pregnancy status. An important initial next step is to define the temporal relevance of the variables described. One challenging element of identifying pregnancy status is its transient nature. Recognizing a

patient was pregnant two years ago has very little baring on her current pregnancy status. In like manner, a positive pregnancy test from many months ago may be meaningless for the patient today. Each of the variables described in this dissertation will need to have a temporal window associated with it. From the date of the finding, how far along can the pregnancy be and how much longer would the pregnancy be expected to last? In this way, rather than a single point in time, each variable will have a window wherein pregnancy is more likely.

Multiple Logistic Regression Modelling

With the variables defined temporally, a multiple logistic modelling approach may be taken to create a predictive model. The patients identified in this dissertation with well-defined temporal bounds on pregnancy episodes may be randomly divided into a training and a test group. Age-matched non-pregnant controls may also be included. Further chart review may be required to ensure pregnancy status of both groups. Sampling may occur at various time points during the year of interest. At each time point, the pregnancy status of each patient will be identified and the predictive model computed. With a model in place, it can be evaluated using the test population.

Evaluation and Validation at Novel Sites

With a predictive model developed at UNMC, efforts may be made to replicate and validate this model at additional sites. The first step in this process will be to survey other sites to ascertain what proportion of the variables identified are commonly available. ETLs for obtaining these data may also be shared as necessary. If sites are able to extract the necessary data elements, the predictive model may be shared and assessed against novel patient populations. IRB approval for chart reviews at each site may be necessary to validate the results. The data survey and model testing may be

done across the GPC where federated queries are supported. To enable testing this model within PCORnet, an evaluation of the variables in light of the CDM would be necessary. If the CDM has enough data variety to encompass all elements of the model, federated queries could be generated and shared with participating sites. Validation across many sites would allow for refinement of the model and development of a shareable computable phenotype suitable for consistent incorporation into future studies.

Future Research Incorporating Location-based Data into the i2b2 Infrastructure

Incorporating Additional Variables in a Standardized Fashion

As described in this dissertation, some risk models incorporate data extracted from sources beyond the EHR. I demonstrated a means of coupling ACS data with clinical data within a de-identified i2b2 infrastructure. Further work is required to instantiate additional variables in a standardized fashion. The approach of extraction of data, linking to patient data, and de-identification may be applied to any single ACS variable or combination of variables to compute indices for socioeconomic status. To facilitate interoperability a standard ontology may be developed to identify which ACS variable is being incorporated. While the ACS has an identifiable code for each data broad element (percent unemployment), each variable within the ACS is often reported for different subsets of the population (i.e. percent unemployment for males, percent unemployment for females). For this reason, it is insufficient to identify the variable in the database with only the ACS identifier as this may represent dozens of values. Developing a standard approach is further hampered as ACS values may be used to compute novel values more meaningful for research. Additionally, as demonstrated herein, these computed values may be combined to calculate indices or standardized for a population of interest. Input from terminologists, epidemiologists, and clinical

researchers may guide the development of a standard ontology for incorporating ACS variables into a clinical data research warehouse.

Linking Address Changes with Appropriate Socioeconomic Data

Along with a standard approach to incorporating additional ACS variables, developing an approach to deal with address changes and updated ACS data is imperative. The ACS is updated on an annual basis and reports data for a five-year period at the block group level. Patients may move one or more times within the catchment area of an academic medical center. The relation between socioeconomic variables and clinical outcomes may be identified more accurately if patients are assigned to the appropriate date range of data for their location during that period. This is a two-fold effort. First, a patient's location may be stored as a fact with a start date and an end date. Any changes in location should be mapped to a new block group with potential changes in socioeconomic status variables. Next, the temporal nature of a patient's location should be linked to the appropriate ACS data set. This is important as areas change over time and as block group boundaries change over time. Mapping a patient's location to the most accurate data will facilitate drawing meaningful clinical conclusions.

Utilizing Included Variables and Indices for Research

As the geocoding process becomes more refined and as additional variables and indices from the ACS are incorporated with clinical data, clinical researchers will have access to extra-EHR data shown to have a major impact on health outcomes. Development of novel as well as refinement of existing socioeconomic indices will be facilitated. A better understanding of elements impacting a patient's clinical outcomes may be developed. With the ability to query a plethora of variables in a de-identified

database, the variables with the greatest clinical relevance may be identified. These data may direct data collection within the clinical environment to describe a patients socioeconomic status and the potential for interventions. In this way, data on a patient's neighborhood extracted from the census may be supplemented by a subset of data offered by the patient.

Conclusions

The data and methods presented in this dissertation represent a step toward facilitating replication in clinical research. I have presented both an approach for evaluating a study and a data set for replication efforts as well as laid the groundwork for filling in commonly missing elements. Further research is needed in many areas to build on the conclusions drawn and to fully develop a computable phenotype for pregnancy status. The approach to replicability evaluation as well as the pipeline for incorporating socioeconomic data into a de-identified data warehouse may be shared and employed at novel sites and across distributed research networks. With the incorporation of extra-EHR data in a queryable fashion along with an approach to using EHR data for replication of risk models and observational studies, additional studies may be validated and extended to novel populations. In this way, the data driving clinical guidelines and practices may be used with greater confidence, hopefully leading to improved patient care.

BIBLIOGRAPHY

1. Meldrum ML. A brief history of the randomized controlled trial. from oranges and lemons to the gold standard. Hematol Oncol Clin North Am 2000;14:745-60, vii.

2. Collins FS, Hudson KL, Briggs JP, Lauer MS. PCORnet: Turning a dream into reality. J Am Med Inform Assoc 2014;21:576-577.

3. Madigan D, Ryan PB, Schuemie M, et al. Evaluating the impact of database heterogeneity on observational study results. Am J Epidemiol 2013;178:645-651.

4. Penberthy LT, Dahman BA, Petkov VI, DeShazo JP. Effort required in eligibility screening for clinical trials. J Oncol Pract 2012;8:365-370.

5. Berndt ER, Cockburn IM. Price indexes for clinical trial research: A feasibility study. Monthly Labor Review 2014;June:11/17/2014. Available at http://www.bls.gov/opub/mlr/2014/article/price-indexes-for-clinical-trial-research-a-feasibility-study.htm .

6. Xu H, Aldrich MC, Chen Q, et al. Validating drug repurposing signals using electronic health records: A case study of metformin associated with reduced cancer mortality. Journal of the American Medical Informatics Association 2014.

7. Emanuel EJ, Schnipper LE, Kamin DY, Levinson J, Lichter AS. The costs of conducting clinical research. J Clin Oncol 2003;21:4145-4150.

8. Califf RM, Zarin DA, Kramer JM, Sherman RE, Aberle LH, Tasneem A. Characteristics of clinical trials registered in ClinicalTrials.gov, 2007-2010. JAMA 2012;307:1838-1847.

9. Bernardez-Pereira S, Lopes RD, Carrion MJ, et al. Prevalence, characteristics, and predictors of early termination of cardiovascular clinical trials due to low recruitment: Insights from the ClinicalTrials.gov registry. Am Heart J 2014;168:213-9.e1.

10. Somkin CP, Altschuler A, Ackerson L, et al. Organizational barriers to physician participation in cancer clinical trials. Am J Manag Care 2005;11:413-421.

11. Christian MC, Trimble EL. Increasing participation of physicians and patients from underrepresented racial and ethnic groups in national cancer institute-sponsored clinical trials. Cancer Epidemiol Biomarkers Prev 2003;12:277s-283s.

12. Food and Drug Administration. Food and Drug Administration Amendments Act of 2007 (FDAAA) 2011.

13. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;3:160018.

14. Riveros C, Dechartres A, Perrodeau E, Haneef R, Boutron I, Ravaud P. Timing and completeness of trial results posted at ClinicalTrials.gov and published in journals. PLoS Med 2013;10:e1001566; discussion e1001566.

15. Bhatt A. Evolution of clinical research: A history before and beyond james lind. Perspect Clin Res 2010;1:6-10.

16. Open Science Collaboration. PSYCHOLOGY. estimating the reproducibility of psychological science. Science 2015;349:aac4716.

17. Vezyridis P, Timmons S. Evolution of primary care databases in UK: A scientometric analysis of research output. BMJ Open 2016;6:e012785-2016-012785.

18. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. J Am Med Inform Assoc 2017;24:198-208.

19. Overhage JM, Ryan PB, Schuemie MJ, Stang PE. Desideratum for evidence based epidemiology. Drug Saf 2013;36:S5-S14.

20. Varas-Lorenzo C, Garcia-Rodriguez LA, Perez-Gutthann S, Duque-Oliart A. Hormone replacement therapy and incidence of acute myocardial infarction. A population-based nested case-control study. Circulation 2000;101:2572-2578.

21. Grodstein F, Clarkson TB, Manson JE. Understanding the divergent data on postmenopausal hormone therapy. N Engl J Med 2003;348:645-650.

22. Cardwell CR, Abnet CC, Cantwell MM, Murray LJ. Exposure to oral bisphosphonates and risk of esophageal cancer. JAMA 2010;304:657-663.

23. Green J, Czanner G, Reeves G, Watson J, Wise L, Beral V. Oral bisphosphonates and risk of cancer of oesophagus, stomach, and colorectum: Case-control analysis within a UK primary care cohort. BMJ 2010;341:c4444.

24. Meier CR, Schlienger RG, Kraenzlin ME, Schlegel B, Jick H. HMG-CoA reductase inhibitors and the risk of fractures. JAMA 2000;283:3205-3210.

25. Lalmohamed A, van Staa TP, Vestergaard P, et al. Statins and risk of lower limb revision surgery: The influence of differences in study design using electronic health records from the united kingdom and denmark. Am J Epidemiol 2016;184:58-66.

26. de Vries F, de Vries C, Cooper C, Leufkens B, van Staa TP. Reanalysis of two studies with contrasting results on the association between statin use and fracture risk: The general practice research database. Int J Epidemiol 2006;35:1301-1308.

27. Wagenmakers EJ, Wetzels R, Borsboom D, van der Maas HL, Kievit RA. An agenda for purely confirmatory research. Perspect Psychol Sci 2012;7:632-638.

28. Simons DJ. The value of direct replication. Perspect Psychol Sci 2014;9:76-80.

29. Cacioppo JT, Kaplan RM, Krosnick JA, Olds JL, Dean H. Social, behavioral, and economic sciences perspectives on robust and reliable science. 2015.

30. Goodman SN, Fanelli D, Ioannidis JP. What does research reproducibility mean? Sci Transl Med 2016;8:341ps12.

31. Begley CG, Ioannidis JP. Reproducibility in science: Improving the standard for basic and preclinical research. Circ Res 2015;116:116-126.

32. Prinz F, Schlange T, Asadullah K. Believe it or not: How much can we rely on published data on potential drug targets? Nature reviews Drug discovery 2011;10:712.

33. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature 2012;483:531.

34. Davison A. Automated capture of experiment context for easier reproducibility in computational research. Computing in Science & Engineering 2012;14:48-56.

35. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. PLoS computational biology 2013;9:e1003285.

36. Mesirov JP. Computer science. accessible reproducible research. Science 2010;327:415-416.

37. Peng RD. Reproducible research in computational science. Science 2011;334:1226-1227.

38. Tan TW, Tong JC, Khan AM, de Silva M, Lim KS, Ranganathan S. Advancing standards for bioinformatics activities: Persistence, reproducibility, disambiguation and minimum information about a bioinformatics investigation (MIABi). BMC Genomics 2010;11:S27.

39. Peng RD. Reproducible research and biostatistics. Biostatistics 2009;10:405-408.

40. Peng RD, Dominici F, Zeger SL. Reproducible epidemiologic research. Am J Epidemiol 2006;163:783-789.

41. Lo B. Sharing clinical trial data: Maximizing benefits, minimizing risk. JAMA 2015;313:793-794.

42. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. Nature 2014;505:612-613.

43. Denaxas S, Direk K, Gonzalez-Izquierdo A, et al. Methods for enhancing the reproducibility of biomedical research findings using electronic health records. BioData mining 2017;10:31.

44. Claerbou JF, Karrenfach M. Electronic documents give reproducible research a new meaning. Society of Exploration Geophysicists, 1992.

45. Schroter S, Black N, Evans S, Godlee F, Osorio L, Smith R. What errors do peer reviewers detect, and does training improve their ability to detect them? J R Soc Med 2008;101:507-514.

46. McDonald RJ, Cloft HJ, Kallmes DF. Fate of submitted manuscripts rejected from the american journal of neuroradiology: Outcomes and commentary. AJNR Am J Neuroradiol 2007;28:1430-1434.

47. Nemery B. What happens to the manuscripts that have not been accepted for publication in occupational and environmental medicine? Occup Environ Med 2001;58:604-607.

48. US Department of Health and Human Services. Office of research integrity annual report 2011. US HHS 2011.

49. Dwan K, Gamble C, Williamson PR, Kirkham JJ, Reporting Bias Group. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. PLoS One 2013;8:e66844.

50. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. Lancet 1991;337:867-872.

51. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. Onkologie 2000;23:597-602.

52. Franco A, Malhotra N, Simonovits G. Social science. publication bias in the social sciences: Unlocking the file drawer. Science 2014;345:1502-1505.

53. Kivimaki M, Batty GD, Kawachi I, Virtanen M, Singh-Manoux A, Brunner EJ. Don't let the truth get in the way of a good story: An illustration of citation bias in epidemiologic research. Am J Epidemiol 2014;180:446-448.

54. Rosenthal R. The file drawer problem and tolerance for null results. Psychol Bull 1979;86:638.

55. Kerr NL. HARKing: Hypothesizing after the results are known. Personality and Social Psychology Review 1998;2:196-217.

56. Simonsohn U, Nelson LD, Simmons JP. P-curve: A key to the file-drawer. J Exp Psychol : Gen 2014;143:534.

57. Hurlbert SH. Pseudoreplication and the design of ecological field experiments. Ecol Monogr 1984;54:187-211.

58. Rothman KJ. Significance questing. Ann Intern Med 1986;105:445-447.

59. Mills JL. Data torturing 1993.

60. White H. A reality check for data snooping. Econometrica 2000;68:1097-1126.

61. Chan A, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. JAMA 2004;291:2457-2465.

62. Berry D. Multiplicities in cancer research: Ubiquitous and necessary evils. J Natl Cancer Inst 2012;104:1125-1133.

63. Leamer EE. Specification Searches: Ad Hoc Inference with Nonexperimental Data. John Wiley & Sons Incorporated, 1978.

64. Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? BMJ 2001;322:226-231.

65. Newton KM, Peissig PL, Kho AN, et al. Validation of electronic medical record-based phenotyping algorithms: Results and lessons learned from the eMERGE network. Journal of the American Medical Informatics Association 2013;20:e147-e154.
66. Springate DA, Kontopantelis E, Ashcroft DM, et al. ClinicalCodes: An online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. PloS one 2014;9:e99825.
67. Papez V, Denaxas S, Hemingway H. Evaluation of semantic web technologies for storing computable definitions of electronic health records phenotyping algorithms. arXiv preprint arXiv:1707.07673 2017.
68. Iqbal SA, Wallach JD, Khoury MJ, Schully SD, Ioannidis JP. Reproducible research practices and transparency across the biomedical literature. PLoS biology 2016;14:e1002333.
69. Asendorpf JB, Conner M, De Fruyt F, et al. Recommendations for increasing replicability in psychology. European Journal of Personality 2013;27:108-119.
70. Moher D, Jones A, Lepage L, Consort Group. Use of the CONSORT statement and quality of reports of randomized trials: A comparative before-and-after evaluation. JAMA 2001;285:1992-1995.
71. Von Elm E, Altman DG, Egger M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. PLoS medicine 2007;4:e296.
72. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies conducted using observational routinely-collected health data (RECORD) statement. PLoS medicine 2015;12:e1001885.

73. Hsiao CJ, Hing E. Use and characteristics of electronic health record systems among office-based physician practices: United states, 2001-2013. NCHS Data Brief 2014;(143):1-8.

74. Hufnagel SP. National electronic health record interoperability chronology. Mil Med 2009;174:35-42.

75. Charles D, Gabriel M, Furukawa M. Adoption of electronic health record systems among U.S. non-federal acute care hospitals: 2008-2013. . Washington D.C.: Office of the National Coordinator for Health Information Technology, 2014.

76. Adler-Milstein J, DesRoches CM, Kralovec P, et al. Electronic health record adoption in US hospitals: Progress continues, but challenges persist. Health Aff (Millwood) 2015;34:2174-2180.

77. DeSalvo KB. Connecting health and care for the nation A shared nationwide interoperability roadmap final version 1.0. National Coordinator for Health Information Technology, 2015.

78. Samwald M, Fehre K, De Bruin J, Adlassnig K. The arden syntax standard for clinical decision support: Experiences and directions. J Biomed Inform 2012;45:711-718.

79. Payne PR, Johnson SB, Starren JB, Tilson HH, Dowdy D. Breaking the translational barriers: The value of integrating biomedical informatics and translational research. J Investig Med 2005;53:192-200.

80. Ni Y, Kennebeck S, Dexheimer JW, et al. Automated clinical trial eligibility prescreening: Increasing the efficiency of patient identification for clinical trials in the emergency department. J Am Med Inform Assoc 2014.

81. Hersh WR, Cimino J, Payne PR, et al. Recommendations for the use of operational electronic health record data in comparative effectiveness research. EGEMS (Wash DC) 2013;1:1018-9214.1018. eCollection 2013.

82. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care 2013;51:S30-7.

83. Brennan L, Watson M, Klaber R, Charles T. The importance of knowing context of hospital episode statistics when reconfiguring the NHS. BMJ: British Medical Journal (Online) 2012;344.

84. Green SM. Congruence of disposition after emergency department intubation in the national hospital ambulatory medical care survey. Ann Emerg Med 2013;61:423-426.e8.

85. Nasir K, Lin Z, Bueno H, et al. Is same-hospital readmission rate a good surrogate for all-hospital readmission rate? Med Care 2010;48:477-481.

86. Bourgeois FC, Olson KL, Mandl KD. Patients treated at multiple acute health care facilities: Quantifying information fragmentation. Arch Intern Med 2010;170:1989-1995.

87. Zhang Z, Sun J. Interval censoring. Stat Methods Med Res 2010;19:53-70.

88. Iezzoni LI. Assessing quality using administrative data. Ann Intern Med 1997;127:666-674.

89. O'malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. Health Serv Res 2005;40:1620-1639.

90. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37:394-403.

91. Bernstam E, Herskovic J, Reeder P, Meric-Bernstam F. Oncology research using electronic medical record data. Journal of Clinical Oncology 2010;28:e16501-e16501.

92. Hersh W. Evaluation of biomedical text-mining systems: Lessons learned from information retrieval. Briefings in bioinformatics 2005;6:344-356.

93. Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. Use of electronic medical records for health outcomes research: A literature review. Medical Care Research and Review 2009;66:611-638.

94. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: Rationale and design for the observational medical outcomes partnership. Ann Intern Med 2010;153:600-606.

95. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the sentinel System—a national resource for evidence development. N Engl J Med 2011;364:498-499.

96. Selby JV, Krumholz HM, Kuntz RE, Collins FS. Network news: Powering clinical research. Sci Transl Med 2013;5:182fs13.

97. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc 2014;21:578-582.

98. Trifiro G, Coloma PM, Rijnbeek PR, et al. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: Why and how? J Intern Med 2014;275:551-561.

99. Rijnbeek PR. Converting to a common data model: What is lost in translation? : Commentary on "fidelity assessment of a clinical practice research datalink conversion to the OMOP common data model". Drug Saf 2014;37:893-896.

100. Waitman LR, Aaronson LS, Nadkarni PM, Connolly DW, Campbell JR. The greater plains collaborative: A PCORnet clinical research data network. J Am Med Inform Assoc 2014;21:637-641.

101. Murphy SN, Mendis M, Hackett K, et al. Architecture of the open-source clinical research chart from informatics for integrating biology and the bedside. AMIA Annu Symp Proc 2007:548-552.

102. Vercellini P, Eskenazi B, Consonni D, et al. Oral contraceptives and risk of endometriosis: A systematic review and meta-analysis. Hum Reprod Update 2011;17:159-170.

103. Kent DM, Nelson J, Dahabreh IJ, Rothwell PM, Altman DG, Hayward RA. Risk and treatment effect heterogeneity: Re-analysis of individual participant data from 32 large clinical trials. Int J Epidemiol 2016.

104. Ioannidis JP, Haidich AB, Lau J. Any casualties in the clash of randomised and observational evidence? BMJ 2001;322:879-880.

105. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19:54-60.

106. Tannen RL, Weiner MG, Xie D. Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: Comparison of database and randomised controlled trial findings. BMJ 2009;338:b81.

107. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000;342:1887-1892.

108. Young SS, Karr A. Deming, data and observational studies. Significance 2011;8:116-116-120.

109. Amarasingham R, Moore BJ, Tabak YP, et al. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Med Care 2010;48:981-988.

110. Jencks SF, Williams MV, Coleman EA. Rehospitalizations among patients in the medicare fee-for-service program. N Engl J Med 2009;360:1418-1428.

111. Leppin AL, Gionfriddo MR, Kessler M, et al. Preventing 30-day hospital readmissions: A systematic review and meta-analysis of randomized trials. JAMA Intern Med 2014;174:1095-1107.

112. Seow H, Phillips CO, Rich MW, Spertus JA, Krumholz HM, Lynn J. Isolation of health services research from practice and policy: The example of chronic heart failure management. J Am Geriatr Soc 2006;54:535-540.

113. Donze J, Lipsitz S, Bates DW, Schnipper JL. Causes and patterns of readmissions in patients with common comorbidities: Retrospective cohort study. BMJ 2013;347:f7171.

114. Singal AG, Rahimi RS, Clark C, et al. An automated model using electronic medical record data identifies patients with cirrhosis at high risk for readmission. Clin Gastroenterol Hepatol 2013;11:1335-1341.e1.

115. Tabak YP, Johannes RS, Silber JH. Using automated clinical data for risk adjustment: Development and validation of six disease-specific mortality predictive models for pay-for-performance. Med Care 2007;45:789-805.

116. Centers for Disease Control and Prevention (CDC). HIPAA privacy rule and public health. guidance from CDC and the U.S. department of health and human services. MMWR Suppl 2003;52:1-17, 19-20.

117. Campbell JR, Campbell WS, Hickman H, Pedersen J, McClay J. Employing complex polyhierarchical ontologies and promoting interoperability of i2b2 data systems. AMIA Annu Symp Proc 2015;2015:359-365.

118. Office of the National Coordinator for Health IT. 2016 interoperability standards advisory. 2016. Available at https://www.healthit.gov/sites/default/files/2016-interoperability-standards-advisory-final-508.pdf .

119. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17:124-130.

120. Keenan PS, Normand SL, Lin Z, et al. An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. Circ Cardiovasc Qual Outcomes 2008;1:29-37.

121. Tabak YP, Sun X, Nunez CM, Johannes RS. Using electronic health record data to develop inpatient mortality predictive model: Acute laboratory risk of mortality score (ALaRMS). J Am Med Inform Assoc 2014;21:455-463.

122. Richesson R, Smerek M. Electronic health records-based phenotyping. Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials 2015.

123. Kuperman GJ, Bobb A, Payne TH, et al. Medication-related clinical decision support in computerized provider order entry systems: A review. J Am Med Inform Assoc 2007;14:29-40.

124. Metzger J, Welebob E, Bates DW, Lipsitz S, Classen DC. Mixed results in the safety performance of computerized physician order entry. Health Aff (Millwood) 2010;29:655-663.

125. MacKenzie SL, Wyatt MC, Schuff R, Tenenbaum JD, Anderson N. Practices and perspectives on building integrated data repositories: Results from a 2010 CTSA survey. Journal of the American Medical Informatics Association 2012;19:e119-e124.

126. Bird CE, Seeman T, Escarce JJ, et al. Neighbourhood socioeconomic status and biological 'wear and tear' in a nationally representative sample of US adults. J Epidemiol Community Health 2010;64:860-865.

127. Casey JA, Pollak J, Glymour MM, Mayeda ER, Hirsch AG, Schwartz BS. Measures of SES for electronic health record-based research. Am J Prev Med 2018;54:430-439.

128. LaVeist T, Pollack K, Thorpe Jr R, Fesahazion R, Gaskin D. Place, not race: Disparities dissipate in southwest baltimore when blacks and whites live under similar conditions. Health Aff 2011;30:1880-1887.

129. Gaskin DJ, Thorpe Jr RJ, McGinty EE, et al. Disparities in diabetes: The nexus of race, poverty, and place. Am J Public Health 2014;104:2147-2155.

130. Diez-Roux AV, Nieto FJ, Muntaner C, et al. Neighborhood environments and coronary heart disease: A multilevel analysis. Am J Epidemiol 1997;146:48-63.

131. LeClere FB, Rogers RG, Peters K. Neighborhood social context and racial differences in women's heart disease mortality. J Health Soc Behav 1998:91-107.

132. Kramer MR, Hogue CR. Is segregation bad for your health? Epidemiol Rev 2009;31:178-194.

133. Kandula NR, Wen M, Jacobs EA, Lauderdale DS. Association between neighborhood context and smoking prevalence among asian americans. Am J Public Health 2009;99:885-892.

134. Kimbro RT. Acculturation in context: Gender, age at migration, neighborhood ethnicity, and health behaviors. Social Science Quarterly 2009;90:1145-1166.

135. White K, Borrell LN. Racial/ethnic neighborhood concentration and self-reported health in new york city. Ethn Dis 2006;16:900-908.

136. Dominici F, Peng RD, Bell ML, et al. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. JAMA 2006;295:1127-1134.

137. Bazemore AW, Cottrell EK, Gold R, et al. “Community vital signs”: Incorporating geocoded social determinants into electronic records to promote patient and population health. Journal of the American Medical Informatics Association 2015;23:407-412.

138. Brokamp C, Wolfe C, Lingren T, Harley J, Ryan P. Decentralized and reproducible geocoding and characterization of community and environmental exposures for multisite studies. Journal of the American Medical Informatics Association 2017.

139. Jacquez GM. A research agenda: Does geocoding positional error matter in health GIS studies? Spatial and spatio-temporal epidemiology 2012;3:7-16.

140. Zandbergen PA. A comparison of address point, parcel and street geocoding techniques. Comput , Environ Urban Syst 2008;32:214-232.

141. Lemke D, Mattauch V, Heidinger O, Hense HW. Who hits the mark? A comparative study of the free geocoding services of google and OpenStreetMap. Gesundheitswesen 2015;77:e160-5.

142. Jacquemin B, Lepeule J, Boudier A, et al. Impact of geocoding methods on associations between long-term exposure to urban air pollution and lung function. Environ Health Perspect 2013;121:1054-1060.

143. Weber GM, Murphy SN, McMurry AJ, et al. The shared health research information network (SHRINE): A prototype federated query tool for clinical data repositories. Journal of the American Medical Informatics Association 2009;16:624-630.

144. Holl S, Plum H. Postgis. GeoInformatics 2009;3:34-36.

145. Braveman P, Egerter S, Williams DR. The social determinants of health: Coming of age. Annu Rev Public Health 2011;32:381-398.

146. McMurry AJ, Murphy SN, MacFadden D, et al. SHRINE: Enabling nationally scalable multi-site disease studies. PloS one 2013;8:e55811.

147. UW Health Innovation Program. Health innovation program. area deprivation index. 2014.

148. Knighton AJ, Savitz L, Belnap T, Stephenson B, VanDerslice J. Introduction of an area deprivation index measuring patient socioeconomic status in an integrated health

system: Implications for population health. EGEMS (Wash DC) 2016;4:1238-9214.1238. eCollection 2016.

149. Singh GK. Area deprivation and widening inequalities in US mortality, 1969-1998. Am J Public Health 2003;93:1137-1143.

150. Messer LC, Laraia BA, Kaufman JS, et al. The development of a standardized neighborhood deprivation index. Journal of Urban Health 2006;83:1041-1062.

151. Butler DC, Petterson S, Phillips RL, Bazemore AW. Measures of social deprivation that predict health care access and need within a rational area of primary care service delivery. Health Serv Res 2013;48:539-559.

152. Cayo MR, Talbot TO. Positional error in automated geocoding of residential addresses. International journal of health geographics 2003;2:10.

153. Zimmerman DL, Rushton G, Armstrong MP, et al. Geocoding Health Data: The use of Geographic Codes in Cancer Prevention and Control, Research and Practice. CRC Press, 2007.

154. Gregorio DI, Cromley E, Mrozinski R, Walsh SJ. Subject loss in spatial analysis of breast cancer. Health Place 1999;5:173-177.

155. Oliver MN, Matthews KA, Siadaty M, Hauck FR, Pickle LW. Geographic bias related to geocoding in epidemiologic studies. International Journal of Health Geographics 2005;4:29.

156. Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW. On the wrong side of the tracts? evaluating the accuracy of geocoding in public health research. Am J Public Health 2001;91:1114-1116.

157. Dearwent SM, Jacobs RR, Halbert JB. Locational uncertainty in georeferencing public health datasets. Journal of Exposure Science and Environmental Epidemiology 2001;11:329.

158. Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL. Positional accuracy of geocoded addresses in epidemiologic research. Epidemiology 2003;14:408-412.

159. Kravets N, Hadden WC. The accuracy of address coding and the effects of coding errors. Health Place 2007;13:293-298.

160. Dubowitz T, Heron M, Bird CE, et al. Neighborhood socioeconomic status and fruit and vegetable intake among whites, blacks, and mexican americans in the united states. Am J Clin Nutr 2008;87:1883-1891.

161. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327:557-560.

APPENDIX A - LOCAL INSTRUCTIONS FOR RUNNING FEDERATED SAS QUERIES FROM PCORNET

SAS Based Queries

1. Create trac ticket:
    - Attach zipped query package
    - Assign to Respond to Queries milestone
    - Paste name of query and description into trac ticket
    - Assign to honest broker to run SAS

2. Load query package into SAS:
    - Download and extract zipped query package
    - Load file structure into DEID /d02/queries/BJG/PopMedNet_Queries
        - This may be done manually in SAS Studio, OR,
        - This may be done with WinSCP

3. Update the SAS script:
    - In SAS Studio, in the query package > sasprograms folder, open the master.sas file
    - d02 folder is popmednet work
    - Update user inputs:
        - DMID = C4
        - SiteID = UN
        - Threshold = 11
        - Attrition Table = Y
        - All data tables in lowercase
        - indata = '/d02/queries/data/' (Most current, approved CDM datasets (Instructions to build SAS datasets)
        - infolder = /d02/queries/BJG/PopMedNet_Queries/QUERY_FOLDER/infolder/
        - drnoc = /d02/queries/BJG/PopMedNet_Queries/QUERY_FOLDER/drnoc/
        - dmlocal = /d02/queries/BJG/PopMedNet_Queries/QUERY_FOLDER/dmlocal/
        - sasmacr= /d02/queries/BJG/PopMedNet_Queries/QUERY_FOLDER/infolder/macros/
        - May be more or less user inputs to define
    - Save changes to master.sas

4. Run query:
    - With master.sas open, hit run button (running man icon at top of screen)
    - When complete, review the log to check for errors or warnings

- Review the workplan and ensure the dmlocal and drnoc folders have all expected files created

5. Share results:
    - Either manually or using WinSCP, extract all files from the drnoc folder for the query
    - Zip files in folder with the query name and attach to the trac ticket
    - Assign trac ticket to jmcclay with next task for data oversight committee review
    - Once data oversight committee has approved the request, results may be submitted via PopMedNet adn the trac ticket closed

APPENDIX B – HEART FAILURE RISK MODEL SAMPLE I2B2 QUERY

```
insert into BlueHeronData.QUERY_GLOBAL_TEMP (encounter_num, patient_num,
panel_count)
with t as (
 select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */  f.encounter_num,
f.patient_num
from BlueHeronData.observation_fact f
where
f.concept_cd IN (select concept_cd from  BlueHeronData.concept_dimension   where
concept_path like '\i2b2\Demographics\Marital Status\%')
group by  f.encounter_num , f.patient_num
 )
select t.encounter_num, t.patient_num, 0 as panel_count  from t
<*>
update BlueHeronData.QUERY_GLOBAL_TEMP set panel_count =1 where
BlueHeronData.QUERY_GLOBAL_TEMP.panel_count =  0 and exists ( select 1 from (
select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */  f.encounter_num,
f.patient_num
from BlueHeronData.observation_fact f
where
f.CONCEPT_CD IN (select CONCEPT_CD from
BlueHeronData.CONCEPT_DIMENSION   where CONCEPT_PATH LIKE
'\i2b2\Procedures\cpt?_codes\99201-99499\99221-99239\99221-99223\%' {ESCAPE '?'}
)
  AND  ( f.start_date >= to_date('01-May-2012 00:00:00','DD-MON-YYYY HH24:MI:SS')
AND f.start_date <= to_date('31-Dec-2016 00:00:00','DD-MON-YYYY HH24:MI:SS') )
group by  f.encounter_num , f.patient_num ) t where
BlueHeronData.QUERY_GLOBAL_TEMP.patient_num = t.patient_num  and
BlueHeronData.QUERY_GLOBAL_TEMP.encounter_num = t.encounter_num   )
<*>
update BlueHeronData.QUERY_GLOBAL_TEMP set panel_count =1 where
BlueHeronData.QUERY_GLOBAL_TEMP.panel_count =  0 and exists ( select 1 from (
select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */  f.encounter_num,
f.patient_num
from BlueHeronData.observation_fact f
where
f.CONCEPT_CD IN (select CONCEPT_CD from
BlueHeronData.CONCEPT_DIMENSION   where CONCEPT_PATH LIKE
'\i2b2\Procedures\cpt?_codes\99201-99499\99221-99239\99231-99239\%' {ESCAPE '?'}
)
  AND  ( f.start_date >= to_date('01-May-2012 00:00:00','DD-MON-YYYY HH24:MI:SS')
AND f.start_date <= to_date('31-Dec-2016 00:00:00','DD-MON-YYYY HH24:MI:SS') )
group by  f.encounter_num , f.patient_num ) t where
BlueHeronData.QUERY_GLOBAL_TEMP.patient_num = t.patient_num  and
BlueHeronData.QUERY_GLOBAL_TEMP.encounter_num = t.encounter_num   )
```

```
<*>
update BlueHeronData.QUERY_GLOBAL_TEMP set panel_count =2 where
BlueHeronData.QUERY_GLOBAL_TEMP.panel_count =  1 and exists ( select 1 from (
select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */ f.patient_num
from BlueHeronData.observation_fact f
where
f.concept_cd IN (select concept_cd from  BlueHeronData.concept_dimension   where
concept_path like
'\i2b2\Diagnoses\A18090800\A8359006\A8359014\A8359777\A8342219\A8345316\%')
  AND  ( f.start_date >= to_date('01-May-2012 00:00:00','DD-MON-YYYY HH24:MI:SS')
AND f.start_date <= to_date('31-Dec-2015 00:00:00','DD-MON-YYYY HH24:MI:SS') )
group by  f.patient_num ) t where
BlueHeronData.QUERY_GLOBAL_TEMP.patient_num = t.patient_num    )
<*>
update BlueHeronData.QUERY_GLOBAL_TEMP set panel_count =2 where
BlueHeronData.QUERY_GLOBAL_TEMP.panel_count =  1 and exists ( select 1 from (
select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */ f.patient_num
from BlueHeronData.observation_fact f
where
f.concept_cd IN (select concept_cd from  BlueHeronData.concept_dimension   where
concept_path like
'\i2b2\Diagnoses\A18090800\A8359006\A8359014\A8359777\A8342219\A8345316\402.
11\%')
  AND  ( f.start_date >= to_date('01-May-2012 00:00:00','DD-MON-YYYY HH24:MI:SS')
AND f.start_date <= to_date('31-Dec-2015 00:00:00','DD-MON-YYYY HH24:MI:SS') )
group by  f.patient_num ) t where
BlueHeronData.QUERY_GLOBAL_TEMP.patient_num = t.patient_num    )
<*>
update BlueHeronData.QUERY_GLOBAL_TEMP set panel_count =2 where
BlueHeronData.QUERY_GLOBAL_TEMP.panel_count =  1 and exists ( select 1 from (
select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */ f.patient_num
from BlueHeronData.observation_fact f
where
f.concept_cd IN (select concept_cd from  BlueHeronData.concept_dimension   where
concept_path like
'\i2b2\Diagnoses\A18090800\A8359006\A8359014\A8359777\A8342219\A8345317\402.
91\%')
  AND  ( f.start_date >= to_date('01-May-2012 00:00:00','DD-MON-YYYY HH24:MI:SS')
AND f.start_date <= to_date('31-Dec-2015 00:00:00','DD-MON-YYYY HH24:MI:SS') )
group by  f.patient_num ) t where
BlueHeronData.QUERY_GLOBAL_TEMP.patient_num = t.patient_num    )
<*>
update BlueHeronData.QUERY_GLOBAL_TEMP set panel_count =2 where
BlueHeronData.QUERY_GLOBAL_TEMP.panel_count =  1 and exists ( select 1 from (
select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */ f.patient_num
from BlueHeronData.observation_fact f
where
f.concept_cd IN (select concept_cd from  BlueHeronData.concept_dimension   where
concept_path like
```

```
'\i2b2\Diagnoses\A18090800\A8359006\A8359014\A8359777\A10863172\A10863171\404.01\%')
  AND  ( f.start_date >= to_date('01-May-2012 00:00:00','DD-MON-YYYY HH24:MI:SS')
AND f.start_date <= to_date('31-Dec-2015 00:00:00','DD-MON-YYYY HH24:MI:SS') )
group by  f.patient_num ) t where
BlueHeronData.QUERY_GLOBAL_TEMP.patient_num = t.patient_num    )
<*>
update BlueHeronData.QUERY_GLOBAL_TEMP set panel_count =2 where
BlueHeronData.QUERY_GLOBAL_TEMP.panel_count =  1 and exists ( select 1 from (
select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */ f.patient_num
from BlueHeronData.observation_fact f
where
f.concept_cd IN (select concept_cd from  BlueHeronData.concept_dimension   where
concept_path like
'\i2b2\Diagnoses\A18090800\A8359006\A8359014\A8359777\A10863172\A10863171\404.03\%')
  AND  ( f.start_date >= to_date('01-May-2012 00:00:00','DD-MON-YYYY HH24:MI:SS')
AND f.start_date <= to_date('31-Dec-2015 00:00:00','DD-MON-YYYY HH24:MI:SS') )
group by  f.patient_num ) t where
BlueHeronData.QUERY_GLOBAL_TEMP.patient_num = t.patient_num    )
<*>
update BlueHeronData.QUERY_GLOBAL_TEMP set panel_count =2 where
BlueHeronData.QUERY_GLOBAL_TEMP.panel_count =  1 and exists ( select 1 from (
select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */ f.patient_num
from BlueHeronData.observation_fact f
where
f.concept_cd IN (select concept_cd from  BlueHeronData.concept_dimension   where
concept_path like
'\i2b2\Diagnoses\A18090800\A8359006\A8359014\A8359777\A10863172\A8345328\404.11\%')
  AND  ( f.start_date >= to_date('01-May-2012 00:00:00','DD-MON-YYYY HH24:MI:SS')
AND f.start_date <= to_date('31-Dec-2015 00:00:00','DD-MON-YYYY HH24:MI:SS') )
group by  f.patient_num ) t where
BlueHeronData.QUERY_GLOBAL_TEMP.patient_num = t.patient_num    )
<*>
update BlueHeronData.QUERY_GLOBAL_TEMP set panel_count =2 where
BlueHeronData.QUERY_GLOBAL_TEMP.panel_count =  1 and exists ( select 1 from (
select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */ f.patient_num
from BlueHeronData.observation_fact f
where
f.concept_cd IN (select concept_cd from  BlueHeronData.concept_dimension   where
concept_path like
'\i2b2\Diagnoses\A18090800\A8359006\A8359014\A8359777\A10863172\A8345328\404.13\%')
  AND  ( f.start_date >= to_date('01-May-2012 00:00:00','DD-MON-YYYY HH24:MI:SS')
AND f.start_date <= to_date('31-Dec-2015 00:00:00','DD-MON-YYYY HH24:MI:SS') )
group by  f.patient_num ) t where
BlueHeronData.QUERY_GLOBAL_TEMP.patient_num = t.patient_num    )
<*>
```

```
update BlueHeronData.QUERY_GLOBAL_TEMP set panel_count =2 where
BlueHeronData.QUERY_GLOBAL_TEMP.panel_count =  1 and exists ( select 1 from (
select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */ f.patient_num
from BlueHeronData.observation_fact f
where
f.concept_cd IN (select concept_cd from  BlueHeronData.concept_dimension   where
concept_path like
'\i2b2\Diagnoses\A18090800\A8359006\A8359014\A8359777\A10863172\A8359867\40
4.91\%')
  AND  ( f.start_date >= to_date('01-May-2012 00:00:00','DD-MON-YYYY HH24:MI:SS')
AND f.start_date <= to_date('31-Dec-2015 00:00:00','DD-MON-YYYY HH24:MI:SS') )
group by  f.patient_num ) t where
BlueHeronData.QUERY_GLOBAL_TEMP.patient_num = t.patient_num    )
<*>
update BlueHeronData.QUERY_GLOBAL_TEMP set panel_count =2 where
BlueHeronData.QUERY_GLOBAL_TEMP.panel_count =  1 and exists ( select 1 from (
select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */ f.patient_num
from BlueHeronData.observation_fact f
where
f.concept_cd IN (select concept_cd from  BlueHeronData.concept_dimension   where
concept_path like
'\i2b2\Diagnoses\A18090800\A8359006\A8359014\A8359777\A10863172\A8359867\40
4.93\%')
  AND  ( f.start_date >= to_date('01-May-2012 00:00:00','DD-MON-YYYY HH24:MI:SS')
AND f.start_date <= to_date('31-Dec-2015 00:00:00','DD-MON-YYYY HH24:MI:SS') )
group by  f.patient_num ) t where
BlueHeronData.QUERY_GLOBAL_TEMP.patient_num = t.patient_num    )
<*>
update BlueHeronData.QUERY_GLOBAL_TEMP set panel_count =2 where
BlueHeronData.QUERY_GLOBAL_TEMP.panel_count =  1 and exists ( select 1 from (
select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */ f.patient_num
from BlueHeronData.observation_fact f
where
f.concept_cd IN (select concept_cd from  BlueHeronData.concept_dimension   where
concept_path like
'\i2b2\Diagnoses\A18090800\A8359006\A8359014\A8360933\A8339688\A19383986\%')
  AND  ( f.start_date >= to_date('01-May-2012 00:00:00','DD-MON-YYYY HH24:MI:SS')
AND f.start_date <= to_date('31-Dec-2015 00:00:00','DD-MON-YYYY HH24:MI:SS') )
group by  f.patient_num ) t where
BlueHeronData.QUERY_GLOBAL_TEMP.patient_num = t.patient_num    )
<*>
update BlueHeronData.QUERY_GLOBAL_TEMP set panel_count =2 where
BlueHeronData.QUERY_GLOBAL_TEMP.panel_count =  1 and exists ( select 1 from (
select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */ f.patient_num
from BlueHeronData.observation_fact f
where
f.concept_cd IN (select concept_cd from  BlueHeronData.concept_dimension   where
concept_path like
'\i2b2\Diagnoses\A18090800\A8359006\A8359014\A8360933\A8339688\425.4\%')
```

```
  AND  ( f.start_date >= to_date('01-May-2012 00:00:00','DD-MON-YYYY HH24:MI:SS')
AND f.start_date <= to_date('31-Dec-2015 00:00:00','DD-MON-YYYY HH24:MI:SS') )
group by  f.patient_num ) t where
BlueHeronData.QUERY_GLOBAL_TEMP.patient_num = t.patient_num    )
<*>
update BlueHeronData.QUERY_GLOBAL_TEMP set panel_count =2 where
BlueHeronData.QUERY_GLOBAL_TEMP.panel_count =  1 and exists ( select 1 from (
select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */ f.patient_num
from BlueHeronData.observation_fact f
where
f.concept_cd IN (select concept_cd from  BlueHeronData.concept_dimension   where
concept_path like
'\i2b2\Diagnoses\A18090800\A8359006\A8359014\A8360933\A8339688\425.5\%')
  AND  ( f.start_date >= to_date('01-May-2012 00:00:00','DD-MON-YYYY HH24:MI:SS')
AND f.start_date <= to_date('31-Dec-2015 00:00:00','DD-MON-YYYY HH24:MI:SS') )
group by  f.patient_num ) t where
BlueHeronData.QUERY_GLOBAL_TEMP.patient_num = t.patient_num    )
<*>
update BlueHeronData.QUERY_GLOBAL_TEMP set panel_count =2 where
BlueHeronData.QUERY_GLOBAL_TEMP.panel_count =  1 and exists ( select 1 from (
select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */ f.patient_num
from BlueHeronData.observation_fact f
where
f.concept_cd IN (select concept_cd from  BlueHeronData.concept_dimension   where
concept_path like
'\i2b2\Diagnoses\A18090800\A8359006\A8359014\A8360933\A8339688\425.7\%')
  AND  ( f.start_date >= to_date('01-May-2012 00:00:00','DD-MON-YYYY HH24:MI:SS')
AND f.start_date <= to_date('31-Dec-2015 00:00:00','DD-MON-YYYY HH24:MI:SS') )
group by  f.patient_num ) t where
BlueHeronData.QUERY_GLOBAL_TEMP.patient_num = t.patient_num    )
<*>
update BlueHeronData.QUERY_GLOBAL_TEMP set panel_count =2 where
BlueHeronData.QUERY_GLOBAL_TEMP.panel_count =  1 and exists ( select 1 from (
select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */ f.patient_num
from BlueHeronData.observation_fact f
where
f.concept_cd IN (select concept_cd from  BlueHeronData.concept_dimension   where
concept_path like
'\i2b2\Diagnoses\A18090800\A8359006\A8359014\A8360933\A8339688\425.8\%')
  AND  ( f.start_date >= to_date('01-May-2012 00:00:00','DD-MON-YYYY HH24:MI:SS')
AND f.start_date <= to_date('31-Dec-2015 00:00:00','DD-MON-YYYY HH24:MI:SS') )
group by  f.patient_num ) t where
BlueHeronData.QUERY_GLOBAL_TEMP.patient_num = t.patient_num    )
<*>
update BlueHeronData.QUERY_GLOBAL_TEMP set panel_count =2 where
BlueHeronData.QUERY_GLOBAL_TEMP.panel_count =  1 and exists ( select 1 from (
select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */ f.patient_num
from BlueHeronData.observation_fact f
where
```

```
f.concept_cd IN (select concept_cd from  BlueHeronData.concept_dimension   where
concept_path like
'\i2b2\Diagnoses\A18090800\A8359006\A8359014\A8360933\A8339688\425.9\%')
  AND  ( f.start_date >= to_date('01-May-2012 00:00:00','DD-MON-YYYY HH24:MI:SS')
AND f.start_date <= to_date('31-Dec-2015 00:00:00','DD-MON-YYYY HH24:MI:SS') )
group by  f.patient_num ) t where
BlueHeronData.QUERY_GLOBAL_TEMP.patient_num = t.patient_num    )
<*>
update BlueHeronData.QUERY_GLOBAL_TEMP set panel_count =2 where
BlueHeronData.QUERY_GLOBAL_TEMP.panel_count =  1 and exists ( select 1 from (
select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */ f.patient_num
from BlueHeronData.observation_fact f
where
f.concept_cd IN (select concept_cd from  BlueHeronData.concept_dimension   where
concept_path like
'\i2b2\Diagnoses\A18090800\A8359006\A8359014\A8360933\A8339687\%')
  AND  ( f.start_date >= to_date('01-May-2012 00:00:00','DD-MON-YYYY HH24:MI:SS')
AND f.start_date <= to_date('31-Dec-2015 00:00:00','DD-MON-YYYY HH24:MI:SS') )
group by  f.patient_num ) t where
BlueHeronData.QUERY_GLOBAL_TEMP.patient_num = t.patient_num    )
<*>
 update BlueHeronData.QUERY_GLOBAL_TEMP set panel_count = -1  where
BlueHeronData.QUERY_GLOBAL_TEMP.panel_count =  2 and exists ( select 1 from (
select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */  f.encounter_num,
f.patient_num
from BlueHeronData.observation_fact f
where
f.concept_cd IN (select concept_cd from  BlueHeronData.concept_dimension   where
concept_path LIKE '\ICD10CM?_2015AA\(Z00-Z99) Fact~rmj9\(Z40-Z53)
Enco~eivf\(Z53) Persons~y7wq\(Z53.2) Proced~esd4\(Z53.21) Proce~0aof\%' {ESCAPE
'?'} )
group by  f.encounter_num , f.patient_num ) t where
BlueHeronData.QUERY_GLOBAL_TEMP.patient_num = t.patient_num  and
BlueHeronData.QUERY_GLOBAL_TEMP.encounter_num = t.encounter_num   )
<*>
 update BlueHeronData.QUERY_GLOBAL_TEMP set panel_count = -1  where
BlueHeronData.QUERY_GLOBAL_TEMP.panel_count =  2 and exists ( select 1 from (
select  /*+ index(observation_fact fact_cnpt_pat_enct_idx) */  f.encounter_num,
f.patient_num
from BlueHeronData.observation_fact f
where
f.concept_cd IN (select concept_cd from  BlueHeronData.concept_dimension   where
concept_path like '\i2b2\DRG\SURG\Orthopedics\BILATERAL OR MULTIPLE MAJOR
JOINT PROCS OF LOWER EXTREMITY W MCC\%')
group by  f.encounter_num , f.patient_num ) t where
BlueHeronData.QUERY_GLOBAL_TEMP.patient_num = t.patient_num  and
BlueHeronData.QUERY_GLOBAL_TEMP.encounter_num = t.encounter_num   )
<*>
```

```
 insert into BlueHeronData.DX (  patient_num  , encounter_num  ) select * from ( select distinct  patient_num  , encounter_num from BlueHeronData.QUERY_GLOBAL_TEMP where panel_count = 2 ) q
```

APPENDIX C – READMISSION RISK MODEL SQL SCRIPT

```
CREATE OR REPLACE PROCEDURE "JGARDNER"."TABAK_SCORING_PROC"

AS

my_date date;
rec_count int;
i int;
sql_string varchar2(32767);
table_name varchar2(100);
lab_mapping_table varchar2(100):='LAB_LOINC_MAPPING'; --change cursor if
changing table name
missing_value_threshold integer:=30; --Number of missing lab values allowed to
have Tabak score calculated
lab_count_variable integer;
percent_of_encs number(10,7);
total_encs integer;

cursor sel_cur is
    select distinct(LAB_LABEL)
    from JGARDNER.LAB_LOINC_MAPPING;
sel_rec sel_cur%ROWTYPE;



BEGIN
select sysdate into my_date from dual;
DBMS_OUTPUT.PUT_LINE('***Start of TABAK_SCORING_PROC');
dbms_output.PUT_LINE('***Creating '||lab_mapping_table||' table at ' ||
my_date);

/*  Ensure table is created prior to running procedure.
--CREATE LAB_LOINC_MAPPING TABLE:
    sql_string := q'[
        CREATE TABLE JGARDNER.]'||lab_mapping_table||q'[
            (   LAB_LABEL VARCHAR2(150),
                LAB_LOINC VARCHAR2(13),
                REFERENCE_LOW NUMBER(6,2),
                REFERENCE_HIGH NUMBER(6,2))
        ]';

    DBMS_OUTPUT.PUT_LINE(sql_string);
    execute immediate sql_string;

    sql_string:=q'[CREATE TABLE JGARDNER.TABAK_SUMMARY
            (LAB_LABEL VARCHAR2(50),
            ENCOUNTERS INTEGER,
            PERCENT_OF_ENCOUNTERS NUMBER(10,7))];
    --DBMS_OUTPUT.PUT_LINE(sql_string);
    execute immediate sql_string;
```

```
    */
--TRUNCATE TABAK_SUMMARY TABLE:
execute immediate 'TRUNCATE TABLE JGARDNER.TABAK_SUMMARY';

--TRUNCATE LAB_MAPPING_TABLE:
execute immediate 'TRUNCATE TABLE JGARDNER.'||lab_mapping_table;

--INSERT VALUES INTO LAB_LOINC_MAPPING TABLE:
    sql_string := q'[INSERT ALL
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('ALBUMIN', 'LOINC:1751-7', 3.3, NULL)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('ALBUMIN', 'LOINC:2862-1', 3.3, NULL)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('AST', 'LOINC:1920-8', NULL, 30)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('AST', 'LOINC:88112-8', NULL, 30)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('AST', 'LOINC:30239-8', NULL, 30)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('TOTAL_BILIRUBIN', 'LOINC:1975-2', NULL, 1.4)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('TOTAL_BILIRUBIN', 'LOINC:42719-5', NULL, 1.4)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('CALCIUM', 'LOINC:17861-6', 8.5, 10.1)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('CALCIUM', 'LOINC:17863-2', 8.5, 10.1)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('CALCIUM','LOINC:17864-0', 8.5, 10.1)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('CALCIUM','LOINC:42567-8', 8.5, 10.1)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('CALCIUM', 'LOINC:57333-7', 8.5, 10.1)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('CALCIUM', 'LOINC:34907-6', 8.5, 10.1)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('CREATININE', 'LOINC:2160-0', NULL, 2)
        INTO JGARDNER.]'||lab_mapping_table||q'[
```

```
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('PRO_BNP', 'LOINC:71425-3', NULL, 8000)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('PRO_BNP', 'LOINC:33762-6', NULL, 8000)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('PRO_BNP', 'LOINC:83107-3', NULL, 8000)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('BNP', 'LOINC:42637-9', NULL, 1200)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('BNP', 'LOINC:30934-4', NULL, 1200)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('GLUCOSE', 'LOINC:2339-0', 71, 135)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('GLUCOSE', 'LOINC:2340-8', 71, 135)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('GLUCOSE', 'LOINC:2345-7', 71, 135)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('GLUCOSE', 'LOINC:14749-6', 71, 135)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('GLUCOSE', 'LOINC:1558-6', 71, 135)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('GLUCOSE', 'LOINC:14771-0', 71, 135)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('GLUCOSE', 'LOINC:15074-8', 71, 135)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('GLUCOSE', 'LOINC:32016-8', 71, 135)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('GLUCOSE', 'LOINC:1556-0', 71, 135)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('GLUCOSE', 'LOINC:27353-2', 71, 135)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('POTASSIUM', 'LOINC:75940-7', 3.3, 4.9)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('POTASSIUM', 'LOINC:22760-3', 3.3, 4.9)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('POTASSIUM', 'LOINC:42569-4', 3.3, 4.9)
    INTO JGARDNER.]'||lab_mapping_table||q'[
```

```
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('POTASSIUM', 'LOINC:77142-8', 3.3, 4.9)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('POTASSIUM', 'LOINC:51618-7', 3.3, 4.9)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('POTASSIUM', 'LOINC:2823-3', 3.3, 4.9)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('POTASSIUM', 'LOINC:6298-4', 3.3, 4.9)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('SODIUM', 'LOINC:2951-2', 136, 143)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('SODIUM','LOINC:2947-0', 136, 143)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('SODIUM','LOINC:42570-2', 136, 143)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('SODIUM','LOINC:77139-4', 136, 143)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('ALKALINE_PHOS','LOINC:1783-0', NULL, 115)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('ALKALINE_PHOS','LOINC:15148-0', NULL, 115)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('ALKALINE_PHOS','LOINC:6768-6', NULL, 115)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('BUN','LOINC:3094-0', NULL, 25)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('BUN','LOINC:6299-2', NULL, 25)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('BUN','LOINC:12964-3', NULL, 25)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('BUN','LOINC:35234-4', NULL, 25)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('PH_ARTERIAL','LOINC:2744-1', 7.36, 7.48)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('PH_ARTERIAL','LOINC:33254-4', 7.36, 7.48)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('PO2_ARTERIAL','LOINC:2703-7', 55.1, 140)
    INTO JGARDNER.]'||lab_mapping_table||q'[
```

```
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('PO2_ARTERIAL','LOINC:19255-9', 55.1, 140)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('PCO2_ARTERIAL','LOINC:2019-8', 36, 50)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('PCO2_ARTERIAL','LOINC:32771-8', 36, 50)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('PTT','LOINC:3173-2', 23, 45)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('PTT','LOINC:14979-9', 23, 45)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('PTT','LOINC:40100-0', 23, 45)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('PT_INR','LOINC:6301-6', NULL, 1.1)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('PT_INR', 'LOINC:34714-6', NULL, 1.1)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('PT_INR', 'LOINC:38875-1', NULL, 1.1)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('BANDS', 'LOINC:26508-2', NULL, 6)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('BANDS', 'LOINC:13354-6', NULL, 6)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('BANDS', 'LOINC:26510-8', NULL, 6)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('BANDS', 'LOINC:764-1', NULL, 6)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('BANDS', 'LOINC:35332-6', NULL, 6)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('HB', 'LOINC:718-7', 11, 18)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('HB', 'LOINC:721-1', 11, 18)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('PLATELETS', 'LOINC:26515-7', 150.1, 420)
    INTO JGARDNER.]'||lab_mapping_table||q'[
        (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
        VALUES ('PLATELETS', 'LOINC:777-3', 150.1, 420)
    INTO JGARDNER.]'||lab_mapping_table||q'[
```

```
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('WBC', 'LOINC:6690-2', 4.4, 10.9)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('WBC', 'LOINC:26464-8', 4.4, 10.9)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('WBC', 'LOINC:751-8', 4.4, 10.9)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('TROPONIN_1', 'LOINC:16255-2', NULL, 0.04)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('TROPONIN_1', 'LOINC:42757-5', NULL, 0.04)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('TROPONIN_1', 'LOINC:10839-9', NULL, 0.04)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('TROPONIN_1', 'LOINC:49563-0', NULL, 0.04)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('CPK_MB', 'LOINC:13969-1', NULL, 2)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('CPK_MB', 'LOINC:83092-7', NULL, 2)
        INTO JGARDNER.]'||lab_mapping_table||q'[
            (LAB_LABEL, LAB_LOINC, REFERENCE_LOW, REFERENCE_HIGH)
            VALUES ('CPK_MB', 'LOINC:49551-5', NULL, 2)
    SELECT * FROM DUAL
    ]';

    --DBMS_OUTPUT.PUT_LINE(sql_string);
    execute immediate sql_string;


--Create potential encounter table:
table_name:='TEMP_POTENTIAL_ENCS';
dbms_output.PUT_LINE('***Creating '||table_name||' table at ' || my_date);
drop_table(table_name);

    sql_string := q'[
    CREATE TABLE JGARDNER.]'||table_name||q'[ AS
        SELECT VD.PATIENT_NUM, VD.ENCOUNTER_NUM, VD.START_DATE,
                VD.DISCHARGE_DISPOSITION, VD.DISCHARGE_STATUS,
                VD.END_DATE - VD.START_DATE AS LOS
            FROM BLUEHERONDATA.VISIT_DIMENSION VD
            WHERE VD.ENC_TYPE IN ('IP','EI')
            AND VD.DISCHARGE_DISPOSITION IS NOT NULL
            AND VD.DISCHARGE_STATUS IS NOT NULL
            AND VD.DISCHARGE_STATUS != 'OT'
            AND VD.DISCHARGE_DISPOSITION IN ('A','E')
            AND VD.END_DATE BETWEEN
                TO_DATE('1/1/2013','mm/dd/yyyy') AND
```

```
                TO_DATE('12/31/2017','mm/dd/yyyy')
            --AND (END_DATE - START_DATE) > 1
    ]';

    --DBMS_OUTPUT.PUT_LINE(sql_string);
    execute immediate sql_string;

    --Count encounters
        sql_string:=q'[select count(*)
            from JGARDNER.TEMP_POTENTIAL_ENCS]';
        --DBMS_OUTPUT.PUT_LINE(sql_string);
        execute immediate sql_string into total_encs;


--Variables:
    --1 Age (at start of encounter encounter)
    --Create potential encounter table:
    table_name:='TEMP_AGE_AT_ENCOUNTER';
    dbms_output.PUT_LINE('***Creating '||table_name||' table at ' || my_date);
    drop_table(table_name);

    sql_string := q'[
        CREATE TABLE JGARDNER.]'||table_name||q'[ AS
            select PE.PATIENT_NUM, PE.ENCOUNTER_NUM,
                round((PE.start_date - PD.birth_date)/365.25,1) AGE_AT_ENC
            FROM BLUEHERONDATA.PATIENT_DIMENSION PD
            JOIN JGARDNER.TEMP_POTENTIAL_ENCS PE
                ON PD.PATIENT_NUM = PE.PATIENT_NUM
        ]';

    --DBMS_OUTPUT.PUT_LINE(sql_string);
    execute immediate sql_string;


    --2 Gender:
    table_name:='TEMP_PATIENT_GENDER';
    dbms_output.PUT_LINE('***Creating '||table_name||' table at ' || my_date);
    drop_table(table_name);

    sql_string := q'[
        CREATE TABLE JGARDNER.]'||table_name||q'[ AS
        SELECT PE.PATIENT_NUM, PD.SEX_CD GENDER
            FROM BLUEHERONDATA.PATIENT_DIMENSION PD
            JOIN JGARDNER.TEMP_POTENTIAL_ENCS PE
                ON PD.PATIENT_NUM = PE.PATIENT_NUM
            WHERE PD.SEX_CD IN ('Female','Male')
            GROUP BY PE.PATIENT_NUM, PD.SEX_CD
        ]';

    --DBMS_OUTPUT.PUT_LINE(sql_string);
    execute immediate sql_string;

    --LAB VALUES:
```

```
open sel_cur;
loop
        fetch sel_cur into sel_rec;
        exit when sel_cur%NOTFOUND; -- no more items, all done
        dbms_output.PUT_LINE('***Creating
JGARDNER.TEMP_'||sel_rec.LAB_LABEL||' table at ' || my_date);
        table_name:='JGARDNER.TEMP_'||sel_rec.LAB_LABEL;
        drop_table(table_name);
        sql_string := q'[CREATE TABLE JGARDNER.TEMP_]'||sel_rec.LAB_LABEL||q'[
as
            SELECT PATIENT_NUM, ENCOUNTER_NUM, START_DATE, CONCEPT_CD,
                    NVAL_NUM, UNITS_CD, LAB_LABEL
                FROM
                    (SELECT PE.PATIENT_NUM, PE.ENCOUNTER_NUM, O.START_DATE,
O.CONCEPT_CD,
                            O.NVAL_NUM, O.UNITS_CD,
']'||sel_rec.LAB_LABEL||q'[' AS LAB_LABEL,
                            ROW_NUMBER() OVER (PARTITION BY O.ENCOUNTER_NUM
                                ORDER BY O.START_DATE ASC) R_NUM
                        FROM JGARDNER.TEMP_POTENTIAL_ENCS PE
                        LEFT JOIN BLUEHERONDATA.OBSERVATION_FACT O
                            ON PE.ENCOUNTER_NUM = O.ENCOUNTER_NUM
                        WHERE PE.LOS > 1 --Eliminate outpatient surgical
procedures
                            and CONCEPT_CD IN (select LAB_LOINC from
JGARDNER.]'||lab_mapping_table||q'[
                            where LAB_LABEL = ']'||sel_rec.LAB_LABEL||q'['))
                WHERE R_NUM = 1
            ]';
        --DBMS_OUTPUT.PUT_LINE(sql_string);
        execute immediate sql_string;

    --Count table
        sql_string:=q'[select count(*)
            from JGARDNER.TEMP_]'||sel_rec.LAB_LABEL;
        --DBMS_OUTPUT.PUT_LINE(sql_string);
        execute immediate sql_string into lab_count_variable;

        sql_string:=q'[select round((]'||to_char(lab_count_variable)||q'[/
                    ]'||to_char(total_encs)||q'[)*100,7)
                    from dual]';
        --DBMS_OUTPUT.PUT_LINE(sql_string);
        execute immediate sql_string into percent_of_encs;

        --Populate summary table
        sql_string := q'[INSERT INTO JGARDNER.TABAK_SUMMARY
                (LAB_LABEL, ENCOUNTERS, PERCENT_OF_ENCOUNTERS)
                VALUES (']'||sel_rec.LAB_LABEL||q'[',
                    ]'||to_char(lab_count_variable)||q'[,
                    ]'||to_char(percent_of_encs)||q'[)
            ]';
        --DBMS_OUTPUT.PUT_LINE(sql_string);
        execute immediate sql_string;
```

```
end loop;
close sel_cur;

    -- raw scoring:
    table_name:='TEMP_TABAK_RAW_SCORE';
    dbms_output.PUT_LINE('***Creating '||table_name||' table at ' || my_date);
    drop_table(table_name);

    sql_string := q'[
        CREATE TABLE JGARDNER.]'||table_name||q'[ AS
        SELECT PE.PATIENT_NUM, PE.ENCOUNTER_NUM, PE.START_DATE,
            PE.DISCHARGE_DISPOSITION, PE.DISCHARGE_STATUS,
            PE.LOS,
            AAE.AGE_AT_ENC,
            PG.GENDER,
            TEMP_ALBUMIN.CONCEPT_CD ALBUMIN_CONCEPT_CD,
                TEMP_ALBUMIN.NVAL_NUM ALBUMIN_NVAL_NUM,
                TEMP_ALBUMIN.UNITS_CD ALBUMIN_UNITS_CD,
            TEMP_AST.CONCEPT_CD AST_CONCEPT_CD,
                TEMP_AST.NVAL_NUM AST_NVAL_NUM,
                TEMP_AST.UNITS_CD AST_UNITS_CD,
            TEMP_TOTAL_BILIRUBIN.CONCEPT_CD TOTAL_BILIRUBIN_CONCEPT_CD,
                TEMP_TOTAL_BILIRUBIN.NVAL_NUM TOTAL_BILIRUBIN_NVAL_NUM,
                TEMP_TOTAL_BILIRUBIN.UNITS_CD TOTAL_BILIRUBIN_UNITS_CD,
            TEMP_CALCIUM.CONCEPT_CD CALCIUM_CONCEPT_CD,
                TEMP_CALCIUM.NVAL_NUM CALCIUM_NVAL_NUM,
                TEMP_CALCIUM.UNITS_CD CALCIUM_UNITS_CD,
            TEMP_CREATININE.CONCEPT_CD CREATININE_CONCEPT_CD,
                TEMP_CREATININE.NVAL_NUM CREATININE_NVAL_NUM,
                TEMP_CREATININE.UNITS_CD CREATININE_UNITS_CD,
            TEMP_PRO_BNP.CONCEPT_CD PRO_BNP_CONCEPT_CD,
                TEMP_PRO_BNP.NVAL_NUM PRO_BNP_NVAL_NUM,
                TEMP_PRO_BNP.UNITS_CD PRO_BNP_UNITS_CD,
            TEMP_BNP.CONCEPT_CD BNP_CONCEPT_CD,
                TEMP_BNP.NVAL_NUM BNP_NVAL_NUM,
                TEMP_BNP.UNITS_CD BNP_UNITS_CD,
            TEMP_GLUCOSE.CONCEPT_CD GLUCOSE_CONCEPT_CD,
                TEMP_GLUCOSE.NVAL_NUM GLUCOSE_NVAL_NUM,
                TEMP_GLUCOSE.UNITS_CD GLUCOSE_UNITS_CD,
            TEMP_POTASSIUM.CONCEPT_CD POTASSIUM_CONCEPT_CD,
                TEMP_POTASSIUM.NVAL_NUM POTASSIUM_NVAL_NUM,
                TEMP_POTASSIUM.UNITS_CD POTASSIUM_UNITS_CD,
            TEMP_SODIUM.CONCEPT_CD SODIUM_CONCEPT_CD,
                TEMP_SODIUM.NVAL_NUM SODIUM_NVAL_NUM,
                TEMP_SODIUM.UNITS_CD SODIUM_UNITS_CD,
            TEMP_ALKALINE_PHOS.CONCEPT_CD ALKALINE_PHOS_CONCEPT_CD,
                TEMP_ALKALINE_PHOS.NVAL_NUM ALKALINE_PHOS_NVAL_NUM,
                TEMP_ALKALINE_PHOS.UNITS_CD ALKALINE_PHOS_UNITS_CD,
            TEMP_BUN.CONCEPT_CD BUN_CONCEPT_CD,
                TEMP_BUN.NVAL_NUM BUN_NVAL_NUM,
                TEMP_BUN.UNITS_CD BUN_UNITS_CD,
            TEMP_PH_ARTERIAL.CONCEPT_CD PH_ARTERIAL_CONCEPT_CD,
                TEMP_PH_ARTERIAL.NVAL_NUM PH_ARTERIAL_NVAL_NUM,
                TEMP_PH_ARTERIAL.UNITS_CD PH_ARTERIAL_UNITS_CD,
```

```
            TEMP_PO2_ARTERIAL.CONCEPT_CD PO2_ARTERIAL_CONCEPT_CD,
                TEMP_PO2_ARTERIAL.NVAL_NUM PO2_ARTERIAL_NVAL_NUM,
                TEMP_PO2_ARTERIAL.UNITS_CD PO2_ARTERIAL_UNITS_CD,
            TEMP_PCO2_ARTERIAL.CONCEPT_CD PCO2_ARTERIAL_CONCEPT_CD,
                TEMP_PCO2_ARTERIAL.NVAL_NUM PCO2_ARTERIAL_NVAL_NUM,
                TEMP_PCO2_ARTERIAL.UNITS_CD PCO2_ARTERIAL_UNITS_CD,
            TEMP_PTT.CONCEPT_CD PTT_CONCEPT_CD,
                TEMP_PTT.NVAL_NUM PTT_NVAL_NUM,
                TEMP_PTT.UNITS_CD PTT_UNITS_CD,
            TEMP_PT_INR.CONCEPT_CD PT_INR_CONCEPT_CD,
                TEMP_PT_INR.NVAL_NUM PT_INR_NVAL_NUM,
                TEMP_PT_INR.UNITS_CD PT_INR_UNITS_CD,
            TEMP_BANDS.CONCEPT_CD BANDS_CONCEPT_CD,
                TEMP_BANDS.NVAL_NUM BANDS_NVAL_NUM,
                TEMP_BANDS.UNITS_CD BANDS_UNITS_CD,
            TEMP_HB.CONCEPT_CD HB_CONCEPT_CD,
                TEMP_HB.NVAL_NUM HB_NVAL_NUM,
                TEMP_HB.UNITS_CD HB_UNITS_CD,
            TEMP_PLATELETS.CONCEPT_CD PLATELETS_CONCEPT_CD,
                TEMP_PLATELETS.NVAL_NUM PLATELETS_NVAL_NUM,
                TEMP_PLATELETS.UNITS_CD PLATELETS_UNITS_CD,
            TEMP_WBC.CONCEPT_CD WBC_CONCEPT_CD,
                TEMP_WBC.NVAL_NUM WBC_NVAL_NUM,
                TEMP_WBC.UNITS_CD WBC_UNITS_CD,
            COALESCE(TEMP_TROPONIN_1.CONCEPT_CD,TEMP_CPK_MB.CONCEPT_CD) AS
TROP_CPK_CONCEPT_CD,
                COALESCE(TEMP_TROPONIN_1.NVAL_NUM,TEMP_CPK_MB.NVAL_NUM) AS
TROP_CPK_NVAL_NUM,
                COALESCE(TEMP_TROPONIN_1.UNITS_CD,TEMP_CPK_MB.UNITS_CD) AS
TROP_CPK_UNITS_CD,
                COALESCE(TEMP_TROPONIN_1.LAB_LABEL,TEMP_CPK_MB.LAB_LABEL) AS
TROP_CPK_LAB_LABEL,
            30 AS TOTAL_LABS,
            999 AS TOTAL_SCORE
        FROM JGARDNER.TEMP_POTENTIAL_ENCS PE
        LEFT JOIN TEMP_ALBUMIN
            ON PE.ENCOUNTER_NUM = TEMP_ALBUMIN.ENCOUNTER_NUM
            AND PE.PATIENT_NUM = TEMP_ALBUMIN.PATIENT_NUM
        LEFT JOIN TEMP_AST
            ON PE.ENCOUNTER_NUM = TEMP_AST.ENCOUNTER_NUM
            AND PE.PATIENT_NUM = TEMP_AST.PATIENT_NUM
        LEFT JOIN TEMP_TOTAL_BILIRUBIN
            ON PE.ENCOUNTER_NUM = TEMP_TOTAL_BILIRUBIN.ENCOUNTER_NUM
            AND PE.PATIENT_NUM = TEMP_TOTAL_BILIRUBIN.PATIENT_NUM
        LEFT JOIN TEMP_CALCIUM
            ON PE.ENCOUNTER_NUM = TEMP_CALCIUM.ENCOUNTER_NUM
            AND PE.PATIENT_NUM = TEMP_CALCIUM.PATIENT_NUM
        LEFT JOIN TEMP_CREATININE
            ON PE.ENCOUNTER_NUM = TEMP_CREATININE.ENCOUNTER_NUM
            AND PE.PATIENT_NUM = TEMP_CREATININE.PATIENT_NUM
        LEFT JOIN TEMP_PRO_BNP
            ON PE.ENCOUNTER_NUM = TEMP_PRO_BNP.ENCOUNTER_NUM
            AND PE.PATIENT_NUM = TEMP_PRO_BNP.PATIENT_NUM
        LEFT JOIN TEMP_BNP
```

```
        ON PE.ENCOUNTER_NUM = TEMP_BNP.ENCOUNTER_NUM
        AND PE.PATIENT_NUM = TEMP_BNP.PATIENT_NUM
    LEFT JOIN TEMP_GLUCOSE
        ON PE.ENCOUNTER_NUM = TEMP_GLUCOSE.ENCOUNTER_NUM
        AND PE.PATIENT_NUM = TEMP_GLUCOSE.PATIENT_NUM
    LEFT JOIN TEMP_POTASSIUM
        ON PE.ENCOUNTER_NUM = TEMP_POTASSIUM.ENCOUNTER_NUM
        AND PE.PATIENT_NUM = TEMP_POTASSIUM.PATIENT_NUM
    LEFT JOIN TEMP_SODIUM
        ON PE.ENCOUNTER_NUM = TEMP_SODIUM.ENCOUNTER_NUM
        AND PE.PATIENT_NUM = TEMP_SODIUM.PATIENT_NUM
    LEFT JOIN TEMP_ALKALINE_PHOS
        ON PE.ENCOUNTER_NUM = TEMP_ALKALINE_PHOS.ENCOUNTER_NUM
        AND PE.PATIENT_NUM = TEMP_ALKALINE_PHOS.PATIENT_NUM
    LEFT JOIN TEMP_BUN
        ON PE.ENCOUNTER_NUM = TEMP_BUN.ENCOUNTER_NUM
        AND PE.PATIENT_NUM = TEMP_BUN.PATIENT_NUM
    LEFT JOIN TEMP_PH_ARTERIAL
        ON PE.ENCOUNTER_NUM = TEMP_PH_ARTERIAL.ENCOUNTER_NUM
        AND PE.PATIENT_NUM = TEMP_PH_ARTERIAL.PATIENT_NUM
    LEFT JOIN TEMP_PO2_ARTERIAL
        ON PE.ENCOUNTER_NUM = TEMP_PO2_ARTERIAL.ENCOUNTER_NUM
        AND PE.PATIENT_NUM = TEMP_PO2_ARTERIAL.PATIENT_NUM
    LEFT JOIN TEMP_PCO2_ARTERIAL
        ON PE.ENCOUNTER_NUM = TEMP_PCO2_ARTERIAL.ENCOUNTER_NUM
        AND PE.PATIENT_NUM = TEMP_PCO2_ARTERIAL.PATIENT_NUM
    LEFT JOIN TEMP_PTT
        ON PE.ENCOUNTER_NUM = TEMP_PTT.ENCOUNTER_NUM
        AND PE.PATIENT_NUM = TEMP_PTT.PATIENT_NUM
    LEFT JOIN TEMP_PT_INR
        ON PE.ENCOUNTER_NUM = TEMP_PT_INR.ENCOUNTER_NUM
        AND PE.PATIENT_NUM = TEMP_PT_INR.PATIENT_NUM
    LEFT JOIN TEMP_BANDS
        ON PE.ENCOUNTER_NUM = TEMP_BANDS.ENCOUNTER_NUM
        AND PE.PATIENT_NUM = TEMP_BANDS.PATIENT_NUM
    LEFT JOIN TEMP_HB
        ON PE.ENCOUNTER_NUM = TEMP_HB.ENCOUNTER_NUM
        AND PE.PATIENT_NUM = TEMP_HB.PATIENT_NUM
    LEFT JOIN TEMP_PLATELETS
        ON PE.ENCOUNTER_NUM = TEMP_PLATELETS.ENCOUNTER_NUM
        AND PE.PATIENT_NUM = TEMP_PLATELETS.PATIENT_NUM
    LEFT JOIN TEMP_WBC
        ON PE.ENCOUNTER_NUM = TEMP_WBC.ENCOUNTER_NUM
        AND PE.PATIENT_NUM = TEMP_WBC.PATIENT_NUM
    LEFT JOIN TEMP_TROPONIN_1
        ON PE.ENCOUNTER_NUM = TEMP_TROPONIN_1.ENCOUNTER_NUM
        AND PE.PATIENT_NUM = TEMP_TROPONIN_1.PATIENT_NUM
    LEFT JOIN TEMP_CPK_MB
        ON PE.ENCOUNTER_NUM = TEMP_CPK_MB.ENCOUNTER_NUM
        AND PE.PATIENT_NUM = TEMP_CPK_MB.PATIENT_NUM
    JOIN TEMP_AGE_AT_ENCOUNTER AAE
        ON PE.PATIENT_NUM = AAE.PATIENT_NUM
        AND PE.ENCOUNTER_NUM = AAE.ENCOUNTER_NUM
    LEFT JOIN TEMP_PATIENT_GENDER PG
```

```
            ON PE.PATIENT_NUM = PG.PATIENT_NUM
        WHERE AAE.AGE_AT_ENC >= 18
        ]';

    --DBMS_OUTPUT.PUT_LINE(sql_string);
    execute immediate sql_string;


--Actual Tabak score
    table_name:='TABAK_SCORE';
    dbms_output.PUT_LINE('***Creating '||table_name||' table at ' || my_date);
    drop_table(table_name);

    sql_string := q'[
        CREATE TABLE JGARDNER.]'||table_name||q'[ AS
        SELECT --RS.PATIENT_NUM,
            RS.ENCOUNTER_NUM, RS.DISCHARGE_DISPOSITION,
            RS.DISCHARGE_STATUS, round(RS.LOS,2) LENGTH_OF_STAY,
            trunc(RS.AGE_AT_ENC) AGE, RS.GENDER,
            999 AS TABAK_TOTAL_SCORE,
            CASE WHEN RS.ALBUMIN_NVAL_NUM IS NULL THEN 1 ELSE 0 END +
                CASE WHEN RS.AST_NVAL_NUM IS NULL THEN 1 ELSE 0 END +
                CASE WHEN RS.TOTAL_BILIRUBIN_NVAL_NUM IS NULL THEN 1 ELSE 0
END +
                CASE WHEN RS.CALCIUM_NVAL_NUM IS NULL THEN 1 ELSE 0 END +
                CASE WHEN RS.CREATININE_NVAL_NUM IS NULL THEN 1 ELSE 0 END +
                CASE WHEN RS.PRO_BNP_NVAL_NUM IS NULL THEN 1 ELSE 0 END +
                CASE WHEN RS.BNP_NVAL_NUM IS NULL THEN 1 ELSE 0 END +
                CASE WHEN RS.GLUCOSE_NVAL_NUM IS NULL THEN 1 ELSE 0 END +
                CASE WHEN RS.POTASSIUM_NVAL_NUM IS NULL THEN 1 ELSE 0 END +
                CASE WHEN RS.SODIUM_NVAL_NUM IS NULL THEN 1 ELSE 0 END +
                CASE WHEN RS.ALKALINE_PHOS_NVAL_NUM IS NULL THEN 1 ELSE 0 END
+
                CASE WHEN RS.BUN_NVAL_NUM IS NULL THEN 1 ELSE 0 END +
                CASE WHEN RS.PH_ARTERIAL_NVAL_NUM IS NULL THEN 1 ELSE 0 END +
                CASE WHEN RS.PO2_ARTERIAL_NVAL_NUM IS NULL THEN 1 ELSE 0 END +
                CASE WHEN RS.PCO2_ARTERIAL_NVAL_NUM IS NULL THEN 1 ELSE 0 END
+
                CASE WHEN RS.PTT_NVAL_NUM IS NULL THEN 1 ELSE 0 END +
                CASE WHEN RS.PT_INR_NVAL_NUM IS NULL THEN 1 ELSE 0 END +
                CASE WHEN RS.BANDS_NVAL_NUM IS NULL THEN 1 ELSE 0 END +
                CASE WHEN RS.HB_NVAL_NUM IS NULL THEN 1 ELSE 0 END +
                CASE WHEN RS.PLATELETS_NVAL_NUM IS NULL THEN 1 ELSE 0 END +
                CASE WHEN RS.WBC_NVAL_NUM IS NULL THEN 1 ELSE 0 END +
                CASE WHEN RS.TROP_CPK_NVAL_NUM IS NULL THEN 1 ELSE 0 END
NumberOfNullFields,
            CASE WHEN RS.AGE_AT_ENC BETWEEN 18 AND 29 THEN 0
                WHEN RS.AGE_AT_ENC BETWEEN 30 AND 34.9 THEN 3
                WHEN RS.AGE_AT_ENC BETWEEN 35 AND 39.9 THEN 10
                WHEN RS.AGE_AT_ENC BETWEEN 40 AND 44.9 THEN 13
                WHEN RS.AGE_AT_ENC BETWEEN 45 AND 49.9 THEN 17
                WHEN RS.AGE_AT_ENC BETWEEN 50 AND 54.9 THEN 20
                WHEN RS.AGE_AT_ENC BETWEEN 55 AND 59.9 THEN 23
                WHEN RS.AGE_AT_ENC BETWEEN 60 AND 64.9 THEN 25
```

```
        WHEN RS.AGE_AT_ENC BETWEEN 65 AND 69.9 THEN 27
        WHEN RS.AGE_AT_ENC BETWEEN 70 AND 74.9 THEN 30
        WHEN RS.AGE_AT_ENC BETWEEN 75 AND 79.9 THEN 32
        WHEN RS.AGE_AT_ENC BETWEEN 80 AND 84.9 THEN 35
        WHEN RS.AGE_AT_ENC BETWEEN 85 AND 89.9 THEN 38
        WHEN RS.AGE_AT_ENC > 89.9 THEN 42
        ELSE 0
        END AGE_SCORE,
    CASE WHEN RS.GENDER = 'Male' THEN 2
        WHEN RS.GENDER = 'Female' THEN 0
        ELSE 0
        END GENDER_SCORE,
    CASE WHEN ALBUMIN_NVAL_NUM <= 2.4 THEN 14
        WHEN ALBUMIN_NVAL_NUM BETWEEN 2.5 AND 2.7 THEN 7
        WHEN ALBUMIN_NVAL_NUM BETWEEN 2.8 AND 3 THEN 4
        WHEN ALBUMIN_NVAL_NUM BETWEEN 3.1 AND 3.3 THEN 1
        WHEN ALBUMIN_NVAL_NUM > 3.3 THEN 0
        ELSE 0
        END AS ALBUMIN_SCORE,
    RS.ALBUMIN_NVAL_NUM,
    CASE WHEN AST_NVAL_NUM <= 30 THEN 0
        WHEN AST_NVAL_NUM BETWEEN 31 AND 40 THEN 2
        WHEN AST_NVAL_NUM BETWEEN 41 AND 60 THEN 4
        WHEN AST_NVAL_NUM BETWEEN 61 AND 100 THEN 6
        WHEN AST_NVAL_NUM > 100 THEN 9
        ELSE 0
        END AS AST_SCORE,
    RS.AST_NVAL_NUM,
    CASE WHEN TOTAL_BILIRUBIN_NVAL_NUM <= 1.4 THEN 0
        WHEN TOTAL_BILIRUBIN_NVAL_NUM BETWEEN 1.5 AND 2 THEN 1
        WHEN TOTAL_BILIRUBIN_NVAL_NUM > 2 THEN 4
        ELSE 0
        END AS TOTAL_BILIRUBIN_SCORE,
    RS.TOTAL_BILIRUBIN_NVAL_NUM,
    CASE WHEN CALCIUM_NVAL_NUM <= 7.9 THEN 4
        WHEN CALCIUM_NVAL_NUM BETWEEN 8 AND 8.4 THEN 1
        WHEN CALCIUM_NVAL_NUM BETWEEN 8.5 AND 10.1 THEN 0
        WHEN CALCIUM_NVAL_NUM > 10.1 THEN 3
        ELSE 0
        END AS CALCIUM_SCORE,
    RS.CALCIUM_NVAL_NUM,
    CASE WHEN CREATININE_NVAL_NUM <=2 THEN 0
        WHEN CREATININE_NVAL_NUM > 2 THEN 1
        ELSE 0
        END AS CREATININE_SCORE,
    RS.CREATININE_NVAL_NUM,
    CASE WHEN PRO_BNP_NVAL_NUM <= 8000 THEN 0
        WHEN PRO_BNP_NVAL_NUM BETWEEN 8001 AND 18000 THEN 5
        WHEN PRO_BNP_NVAL_NUM > 18000 THEN 10
        ELSE 0
        END AS PRO_BNP_SCORE,
    RS.PRO_BNP_NVAL_NUM,
    CASE WHEN BNP_NVAL_NUM <= 1200 THEN 0
        WHEN BNP_NVAL_NUM BETWEEN 1201 AND 2400 THEN 2
```

```
        WHEN BNP_NVAL_NUM > 2400 THEN 4
        ELSE 0
        END AS BNP_SCORE,
    RS.BNP_NVAL_NUM,
    CASE WHEN GLUCOSE_NVAL_NUM <=70 THEN 7
        WHEN GLUCOSE_NVAL_NUM BETWEEN 71 AND 135 THEN 0
        WHEN GLUCOSE_NVAL_NUM BETWEEN 136 AND 165 THEN 2
        WHEN GLUCOSE_NVAL_NUM > 165 THEN 5
        ELSE 0
        END AS GLUCOSE_SCORE,
    RS.GLUCOSE_NVAL_NUM,
    CASE WHEN POTASSIUM_NVAL_NUM <= 3.2 THEN 3
        WHEN POTASSIUM_NVAL_NUM BETWEEN 3.3 AND 4.9 THEN 0
        WHEN POTASSIUM_NVAL_NUM BETWEEN 5 AND 5.3 THEN 2
        WHEN POTASSIUM_NVAL_NUM >5.3 THEN 3
        ELSE 0
        END AS POTASSIUM_SCORE,
    RS.POTASSIUM_NVAL_NUM,
    CASE WHEN SODIUM_NVAL_NUM <= 130    THEN 4
        WHEN SODIUM_NVAL_NUM BETWEEN 131 AND 135 THEN 1
        WHEN SODIUM_NVAL_NUM BETWEEN 136 AND 143 THEN 0
        WHEN SODIUM_NVAL_NUM BETWEEN 144 AND 145 THEN 4
        WHEN SODIUM_NVAL_NUM > 145 THEN 9
        ELSE 0
        END AS SODIUM_SCORE,
    RS.SODIUM_NVAL_NUM,
    CASE WHEN ALKALINE_PHOS_NVAL_NUM <= 115 THEN 0
        WHEN ALKALINE_PHOS_NVAL_NUM BETWEEN 116 AND 220 THEN 2
        WHEN ALKALINE_PHOS_NVAL_NUM BETWEEN 221 AND 630 THEN 5
        WHEN ALKALINE_PHOS_NVAL_NUM >630 THEN 8
        ELSE 0
        END AS ALKALINE_PHOS_SCORE,
    RS.ALKALINE_PHOS_NVAL_NUM,
    CASE WHEN BUN_NVAL_NUM <= 25 THEN 0
        WHEN BUN_NVAL_NUM BETWEEN 26 AND 30 THEN 4
        WHEN BUN_NVAL_NUM BETWEEN 31 AND 40 THEN 6
        WHEN BUN_NVAL_NUM BETWEEN 41 AND 55 THEN 8
        WHEN BUN_NVAL_NUM >55 THEN 10
        ELSE 0
        END AS BUN_SCORE,
    RS.BUN_NVAL_NUM,
    CASE WHEN PH_ARTERIAL_NVAL_NUM <= 7.2 THEN 21
        WHEN PH_ARTERIAL_NVAL_NUM BETWEEN 7.21 AND 7.3 THEN 13
        WHEN PH_ARTERIAL_NVAL_NUM BETWEEN 7.31 AND 7.35 THEN 10
        WHEN PH_ARTERIAL_NVAL_NUM BETWEEN 7.36 AND 7.48 THEN 0
        WHEN PH_ARTERIAL_NVAL_NUM >7.48 THEN 8
        ELSE 0
        END AS PH_ARTERIAL_SCORE,
    RS.PH_ARTERIAL_NVAL_NUM,
    CASE WHEN PO2_ARTERIAL_NVAL_NUM <= 50 THEN 12
        WHEN PO2_ARTERIAL_NVAL_NUM BETWEEN 50.1 AND 55 THEN 9
        WHEN PO2_ARTERIAL_NVAL_NUM BETWEEN 55.1 AND 140 THEN 0
        WHEN PO2_ARTERIAL_NVAL_NUM > 140 THEN 12
        ELSE 0
```

```
    END AS PO2_ARTERIAL_SCORE,
RS.PO2_ARTERIAL_NVAL_NUM,
CASE WHEN PCO2_ARTERIAL_NVAL_NUM <=35 THEN 9
    WHEN PCO2_ARTERIAL_NVAL_NUM BETWEEN 36 AND 50 THEN 0
    WHEN PCO2_ARTERIAL_NVAL_NUM > 50 THEN 7
    ELSE 0
    END AS PCO2_ARTERIAL_SCORE,
RS.PCO2_ARTERIAL_NVAL_NUM,
CASE WHEN PTT_NVAL_NUM <= 22    THEN 3
    WHEN PTT_NVAL_NUM BETWEEN 23 AND 45 THEN 0
    WHEN PTT_NVAL_NUM BETWEEN 45.1 AND 55 THEN 3
    WHEN PTT_NVAL_NUM > 55 THEN 4
    ELSE 0
    END AS PTT_SCORE,
RS.PTT_NVAL_NUM,
CASE WHEN PT_INR_NVAL_NUM <=1.1 THEN 0
    WHEN PT_INR_NVAL_NUM BETWEEN 1.11 AND 1.4 THEN 4
    WHEN PT_INR_NVAL_NUM BETWEEN 1.41 AND 2 THEN 7
    WHEN PT_INR_NVAL_NUM BETWEEN 2.1 AND 5 THEN 5
    WHEN PT_INR_NVAL_NUM > 5 THEN 8
    ELSE 0
    END AS PT_INR_SCORE,
RS.PT_INR_NVAL_NUM,
CASE WHEN BANDS_NVAL_NUM <=6 THEN 0
    WHEN BANDS_NVAL_NUM BETWEEN 7 AND 13 THEN 6
    WHEN BANDS_NVAL_NUM BETWEEN 14 AND 32 THEN 9
    WHEN BANDS_NVAL_NUM > 32 THEN 12
    ELSE 0
    END AS BANDS_SCORE,
RS.BANDS_NVAL_NUM,
CASE WHEN HB_NVAL_NUM <=10 THEN 2
    WHEN HB_NVAL_NUM BETWEEN 11 AND 18 THEN 0
    WHEN HB_NVAL_NUM > 18 THEN 4
    ELSE 0
    END AS HB_SCORE,
RS.HB_NVAL_NUM,
CASE WHEN PLATELETS_NVAL_NUM <=115 THEN 10
    WHEN PLATELETS_NVAL_NUM BETWEEN 115.1 AND 150 THEN 2
    WHEN PLATELETS_NVAL_NUM BETWEEN 150.1 AND 420 THEN 0
    WHEN PLATELETS_NVAL_NUM > 420 THEN 2
    ELSE 0
    END AS PLATELETS_SCORE,
RS.PLATELETS_NVAL_NUM,
CASE WHEN WBC_NVAL_NUM <=4.3 THEN 4
    WHEN WBC_NVAL_NUM BETWEEN 4.4 AND 10.9 THEN 0
    WHEN WBC_NVAL_NUM BETWEEN 11 AND 14.1 THEN 4
    WHEN WBC_NVAL_NUM BETWEEN 14.2 AND 19.8 THEN 7
    WHEN WBC_NVAL_NUM > 19.8 THEN 12
    ELSE 0
    END AS WBC_SCORE,
RS.WBC_NVAL_NUM,
CASE WHEN TROP_CPK_LAB_LABEL = 'TROPONIN_1'
        AND TROP_CPK_NVAL_NUM <= 0.04 THEN 0
    WHEN TROP_CPK_LAB_LABEL = 'TROPONIN_1'
```

```
                AND TROP_CPK_NVAL_NUM BETWEEN 0.05 AND 0.1 THEN 2
            WHEN TROP_CPK_LAB_LABEL = 'TROPONIN_1'
                AND TROP_CPK_NVAL_NUM BETWEEN 0.11 AND 0.2 THEN 4
            WHEN TROP_CPK_LAB_LABEL = 'TROPONIN_1'
                AND TROP_CPK_NVAL_NUM BETWEEN 0.21 AND 0.3 THEN 8
            WHEN TROP_CPK_LAB_LABEL = 'TROPONIN_1'
                AND TROP_CPK_NVAL_NUM >0.3 THEN 13
            WHEN TROP_CPK_LAB_LABEL = 'CPK_MB'
                AND TROP_CPK_NVAL_NUM <= 2 THEN 0
            WHEN TROP_CPK_LAB_LABEL = 'CPK_MB'
                AND TROP_CPK_NVAL_NUM BETWEEN 3 AND 5 THEN 2
            WHEN TROP_CPK_LAB_LABEL = 'CPK_MB'
                AND TROP_CPK_NVAL_NUM BETWEEN 6 AND 10 THEN 4
            WHEN TROP_CPK_LAB_LABEL = 'CPK_MB'
                AND TROP_CPK_NVAL_NUM BETWEEN 11 AND 34 THEN 8
            WHEN TROP_CPK_LAB_LABEL = 'CPK_MB'
                AND TROP_CPK_NVAL_NUM > 34 THEN 13
            ELSE 0
            END AS TROP_CPK_SCORE,
        RS.TROP_CPK_NVAL_NUM,
        RS.TROP_CPK_LAB_LABEL
    FROM TEMP_TABAK_RAW_SCORE RS
]';

--DBMS_OUTPUT.PUT_LINE(sql_string);
execute immediate sql_string;
DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted.');

sql_string:=q'[UPDATE JGARDNER.TABAK_SCORE
    SET TABAK_TOTAL_SCORE = ALBUMIN_SCORE +
    AST_SCORE +
    TOTAL_BILIRUBIN_SCORE +
    CALCIUM_SCORE +
    CREATININE_SCORE +
    PRO_BNP_SCORE +
    BNP_SCORE +
    GLUCOSE_SCORE +
    POTASSIUM_SCORE +
    SODIUM_SCORE +
    ALKALINE_PHOS_SCORE +
    BUN_SCORE +
    PH_ARTERIAL_SCORE +
    PO2_ARTERIAL_SCORE +
    PCO2_ARTERIAL_SCORE +
    PTT_SCORE +
    PT_INR_SCORE +
    BANDS_SCORE +
    HB_SCORE +
    PLATELETS_SCORE +
    WBC_SCORE +
    TROP_CPK_SCORE +
    GENDER_SCORE +
    AGE_SCORE
```

```
    ]';

    --DBMS_OUTPUT.PUT_LINE(sql_string);
    execute immediate sql_string;



--Add columns for binning with 1 = yes and 0 = no:
table_name:='TABAK_SCORE_BINS';
    dbms_output.PUT_LINE('***Creating '||table_name||' table at ' || my_date);
    drop_table(table_name);

    sql_string := q'[
        CREATE TABLE JGARDNER.]'||table_name||q'[ AS
        SELECT TS.*,
            CASE WHEN AGE BETWEEN 18 AND 29.9 THEN 1 ELSE 0 END AS AGE_LT_30,
            CASE WHEN AGE BETWEEN 30 AND 34.9 THEN 1 ELSE 0 END AS AGE_30_34,
            CASE WHEN AGE BETWEEN 35 AND 39.9 THEN 1 ELSE 0 END AS AGE_35_39,
            CASE WHEN AGE BETWEEN 40 AND 44.9 THEN 1 ELSE 0 END AS AGE_40_44,
            CASE WHEN AGE BETWEEN 45 AND 49.9 THEN 1 ELSE 0 END AS AGE_45_49,
            CASE WHEN AGE BETWEEN 50 AND 54.9 THEN 1 ELSE 0 END AS AGE_50_54,
            CASE WHEN AGE BETWEEN 55 AND 59.9 THEN 1 ELSE 0 END AS AGE_55_59,
            CASE WHEN AGE BETWEEN 60 AND 64.9 THEN 1 ELSE 0 END AS AGE_60_64,
            CASE WHEN AGE BETWEEN 65 AND 69.9 THEN 1 ELSE 0 END AS AGE_65_69,
            CASE WHEN AGE BETWEEN 70 AND 74.9 THEN 1 ELSE 0 END AS AGE_70_74,
            CASE WHEN AGE BETWEEN 75 AND 79.9 THEN 1 ELSE 0 END AS AGE_75_79,
            CASE WHEN AGE BETWEEN 80 AND 84.9 THEN 1 ELSE 0 END AS AGE_80_84,
            CASE WHEN AGE BETWEEN 85 AND 89.9 THEN 1 ELSE 0 END AS AGE_85_89,
            CASE WHEN AGE > 89 THEN 1 ELSE 0 END AS AGE_GT_89,
            CASE WHEN GENDER = 'Male' THEN 1 ELSE 0 END AS GENDER_M,
            CASE WHEN GENDER = 'Female' THEN 1 ELSE 0 END AS GENDER_F,
            CASE WHEN ALBUMIN_NVAL_NUM <= 2.4 THEN 1 ELSE 0 END AS
ALBUMIN_LE_24,
            CASE WHEN ALBUMIN_NVAL_NUM BETWEEN 2.5 AND 2.7 THEN 1 ELSE 0 END
AS ALBUMIN_25_27,
            CASE WHEN ALBUMIN_NVAL_NUM BETWEEN 2.8 AND 3 THEN 1 ELSE 0 END AS
ALBUMIN_28_3,
            CASE WHEN ALBUMIN_NVAL_NUM BETWEEN 3.1 AND 3.3 THEN 1 ELSE 0 END
AS ALBUMIN_31_33,
            CASE WHEN ALBUMIN_NVAL_NUM > 3.3 THEN 1 ELSE 0 END AS
ALBUMIN_GT_33,
            CASE WHEN AST_NVAL_NUM <= 30 THEN 1 ELSE 0 END AS AST_LE_30,
            CASE WHEN AST_NVAL_NUM BETWEEN 31 AND 40 THEN 1 ELSE 0 END AS
AST_31_40,
            CASE WHEN AST_NVAL_NUM BETWEEN 41 AND 60 THEN 1 ELSE 0 END AS
AST_41_60,
            CASE WHEN AST_NVAL_NUM BETWEEN 61 AND 100 THEN 1 ELSE 0 END AS
AST_61_100,
            CASE WHEN AST_NVAL_NUM > 100 THEN 1 ELSE 0 END AS AST_GT_100,
            CASE WHEN TOTAL_BILIRUBIN_NVAL_NUM <= 1.4 THEN 1 ELSE 0 END AS
TOT_BIL_LE_14,
            CASE WHEN TOTAL_BILIRUBIN_NVAL_NUM BETWEEN 1.5 AND 2 THEN 1 ELSE 0
END AS TOT_BIL_15_2,
```

```
            CASE WHEN TOTAL_BILIRUBIN_NVAL_NUM > 2 THEN 1 ELSE 0 END AS
TOT_BIL_GT_2,
            CASE WHEN CALCIUM_NVAL_NUM <= 7.9 THEN 1 ELSE 0 END AS
CALCIUM_LE_79,
            CASE WHEN CALCIUM_NVAL_NUM BETWEEN 8 AND 8.4 THEN 1 ELSE 0 END AS
CALCIUM_8_84,
            CASE WHEN CALCIUM_NVAL_NUM BETWEEN 8.5 AND 10.1 THEN 1 ELSE 0 END
AS CALCIUM_85_101,
            CASE WHEN CALCIUM_NVAL_NUM > 10.1 THEN 1 ELSE 0 END AS
CALCIUM_GT_101,
            CASE WHEN CREATININE_NVAL_NUM <=2 THEN 1 ELSE 0 END AS
CREATININE_LE_2,
            CASE WHEN CREATININE_NVAL_NUM > 2 THEN 1 ELSE 0 END AS
CREATININE_GT_2,
            CASE WHEN PRO_BNP_NVAL_NUM <= 8000 THEN 1 ELSE 0 END AS
PRO_BNP_LE_8000,
            CASE WHEN PRO_BNP_NVAL_NUM BETWEEN 8001 AND 18000 THEN 1 ELSE 0
END AS PRO_BNP_8_18,
            CASE WHEN PRO_BNP_NVAL_NUM > 18000 THEN 1 ELSE 0 END AS
PRO_BNP_GT_18,
            CASE WHEN BNP_NVAL_NUM <= 1200 THEN 1 ELSE 0 END AS BNP_LE12,
            CASE WHEN BNP_NVAL_NUM BETWEEN 1201 AND 2400 THEN 1 ELSE 0 END AS
BNP_12_24,
            CASE WHEN BNP_NVAL_NUM > 2400 THEN 1 ELSE 0 END AS BNP_GT24,
            CASE WHEN GLUCOSE_NVAL_NUM <=70 THEN 1 ELSE 0 END AS
GLUCOSE_LE_70,
            CASE WHEN GLUCOSE_NVAL_NUM BETWEEN 71 AND 135 THEN 1 ELSE 0 END AS
GLUCOSE_71_135,
            CASE WHEN GLUCOSE_NVAL_NUM BETWEEN 136 AND 165 THEN 1 ELSE 0 END
AS GLUCOSE_136_165,
            CASE WHEN GLUCOSE_NVAL_NUM > 165 THEN 1 ELSE 0 END AS
GLUCOSE_GT_165,
            CASE WHEN POTASSIUM_NVAL_NUM <= 3.2 THEN 1 ELSE 0 END AS
POTASSIUM_LE_32,
            CASE WHEN POTASSIUM_NVAL_NUM BETWEEN 3.3 AND 4.9 THEN 1 ELSE 0 END
AS POTASSIUM_33_49,
            CASE WHEN POTASSIUM_NVAL_NUM BETWEEN 5 AND 5.3 THEN 1 ELSE 0 END
AS POTASSIUM_5_53,
            CASE WHEN POTASSIUM_NVAL_NUM >5.3 THEN 1 ELSE 0 END AS
POTASSIUM_GT_53,
            CASE WHEN SODIUM_NVAL_NUM <= 130 THEN 1 ELSE 0 END AS
SODIUM_LE_130,
            CASE WHEN SODIUM_NVAL_NUM BETWEEN 131 AND 135 THEN 1 ELSE 0 END AS
SODIUM_131_135,
            CASE WHEN SODIUM_NVAL_NUM BETWEEN 136 AND 143 THEN 1 ELSE 0 END AS
SODIUM_136_143,
            CASE WHEN SODIUM_NVAL_NUM BETWEEN 144 AND 145 THEN 1 ELSE 0 END AS
SODIUM_144_145,
            CASE WHEN SODIUM_NVAL_NUM > 145 THEN 1 ELSE 0 END AS
SODIUM_GT_145,
            CASE WHEN ALKALINE_PHOS_NVAL_NUM <= 115 THEN 1 ELSE 0 END AS
ALKALINE_PHOS_LE_115,
            CASE WHEN ALKALINE_PHOS_NVAL_NUM BETWEEN 116 AND 220 THEN 1 ELSE 0
END AS ALKALINE_PHOS_116_220,
```

```
            CASE WHEN ALKALINE_PHOS_NVAL_NUM BETWEEN 221 AND 630 THEN 1 ELSE 0
END AS ALKALINE_PHOS_221_630,
            CASE WHEN ALKALINE_PHOS_NVAL_NUM >630 THEN 1 ELSE 0 END AS
ALKALINE_PHOS_GT_630,
            CASE WHEN BUN_NVAL_NUM <= 25 THEN 1 ELSE 0 END AS BUN_LE_25,
            CASE WHEN BUN_NVAL_NUM BETWEEN 26 AND 30 THEN 1 ELSE 0 END AS
BUN_26_30,
            CASE WHEN BUN_NVAL_NUM BETWEEN 31 AND 40 THEN 1 ELSE 0 END AS
BUN_31_40,
            CASE WHEN BUN_NVAL_NUM BETWEEN 41 AND 55 THEN 1 ELSE 0 END AS
BUN_41_55,
            CASE WHEN BUN_NVAL_NUM >55 THEN 1 ELSE 0 END AS BUN_GT_55,
            CASE WHEN PH_ARTERIAL_NVAL_NUM <= 7.2 THEN 1 ELSE 0 END AS
PH_ARTERIAL_LE_72,
            CASE WHEN PH_ARTERIAL_NVAL_NUM BETWEEN 7.21 AND 7.3 THEN 1 ELSE 0
END AS PH_ARTERIAL_721_73,
            CASE WHEN PH_ARTERIAL_NVAL_NUM BETWEEN 7.31 AND 7.35 THEN 1 ELSE 0
END AS PH_ARTERIAL_731_735,
            CASE WHEN PH_ARTERIAL_NVAL_NUM BETWEEN 7.36 AND 7.48 THEN 1 ELSE 0
END AS PH_ARTERIAL_736_748,
            CASE WHEN PH_ARTERIAL_NVAL_NUM >7.48 THEN 1 ELSE 0 END AS
PH_ARTERIAL_GT_748,
            CASE WHEN PO2_ARTERIAL_NVAL_NUM <= 50 THEN 1 ELSE 0 END AS
PO2_ARTERIAL_LE_50,
            CASE WHEN PO2_ARTERIAL_NVAL_NUM BETWEEN 50.1 AND 55 THEN 1 ELSE 0
END AS PO2_ARTERIAL_501_55,
            CASE WHEN PO2_ARTERIAL_NVAL_NUM BETWEEN 55.1 AND 140 THEN 1 ELSE 0
END AS PO2_ARTERIAL_551_140,
            CASE WHEN PO2_ARTERIAL_NVAL_NUM > 140 THEN 1 ELSE 0 END AS
PO2_ARTERIAL_GT_140,
            CASE WHEN PCO2_ARTERIAL_NVAL_NUM <=35 THEN 1 ELSE 0 END AS
PCO2_ARTERIAL_LE_35,
            CASE WHEN PCO2_ARTERIAL_NVAL_NUM BETWEEN 36 AND 50 THEN 1 ELSE 0
END AS PCO2_ARTERIAL_36_50,
            CASE WHEN PCO2_ARTERIAL_NVAL_NUM > 50 THEN 1 ELSE 0 END AS
PCO2_ARTERIAL_GT_50,
            CASE WHEN PTT_NVAL_NUM <= 22 THEN 1 ELSE 0 END AS PTT_LE_22,
            CASE WHEN PTT_NVAL_NUM BETWEEN 23 AND 45 THEN 1 ELSE 0 END AS
PTT_23_45,
            CASE WHEN PTT_NVAL_NUM BETWEEN 45.1 AND 55 THEN 1  ELSE 0 END AS
PTT_451_55,
            CASE WHEN PTT_NVAL_NUM > 55 THEN 1 ELSE 0 END AS PTT_GT_55,
            CASE WHEN PT_INR_NVAL_NUM <=1.1 THEN 1 ELSE 0 END AS PT_INR_LE_11,
            CASE WHEN PT_INR_NVAL_NUM BETWEEN 1.11 AND 1.4 THEN 1 ELSE 0 END
AS PT_INR_111_14,
            CASE WHEN PT_INR_NVAL_NUM BETWEEN 1.41 AND 2 THEN 1 ELSE 0 END AS
PT_INR_141_2,
            CASE WHEN PT_INR_NVAL_NUM BETWEEN 2.1 AND 5 THEN 1 ELSE 0 END AS
PT_INR_21_5,
            CASE WHEN PT_INR_NVAL_NUM > 5 THEN 1 ELSE 0 END AS PT_INR_GT_5,
            CASE WHEN BANDS_NVAL_NUM <=6 THEN 1 ELSE 0 END AS BANDS_LE_6,
            CASE WHEN BANDS_NVAL_NUM BETWEEN 7 AND 13 THEN 1 ELSE 0 END AS
BANDS_7_13,
```

```
            CASE WHEN BANDS_NVAL_NUM BETWEEN 14 AND 32 THEN 1 ELSE 0 END AS
BANDS_14_32,
            CASE WHEN BANDS_NVAL_NUM > 32 THEN 1 ELSE 0 END AS BANDS_GT_32,
            CASE WHEN HB_NVAL_NUM <=10 THEN 1 ELSE 0 END AS HB_LE_10,
            CASE WHEN HB_NVAL_NUM BETWEEN 11 AND 18 THEN 1 ELSE 0 END AS
HB_11_18,
            CASE WHEN HB_NVAL_NUM > 18 THEN 1 ELSE 0 END AS HB_GT_18,
            CASE WHEN PLATELETS_NVAL_NUM <=115 THEN 1 ELSE 0 END AS
PLATELETS_LE_115,
            CASE WHEN PLATELETS_NVAL_NUM BETWEEN 115.1 AND 150 THEN 1 ELSE 0
END AS PLATELETS_115_150,
            CASE WHEN PLATELETS_NVAL_NUM BETWEEN 150.1 AND 420 THEN 1 ELSE 0
END AS PLATELETS_150_420,
            CASE WHEN PLATELETS_NVAL_NUM > 420 THEN 1 ELSE 0 END AS
PLATELETS_GT_420,
            CASE WHEN WBC_NVAL_NUM <=4.3 THEN 1 ELSE 0 END AS WBC_LE_43,
            CASE WHEN WBC_NVAL_NUM BETWEEN 4.4 AND 10.9 THEN 1 ELSE 0 END AS
WBC_44_109,
            CASE WHEN WBC_NVAL_NUM BETWEEN 11 AND 14.1 THEN 1 ELSE 0 END AS
WBC_11_141,
            CASE WHEN WBC_NVAL_NUM BETWEEN 14.2 AND 19.8 THEN 1 ELSE 0 END AS
WBC_142_198,
            CASE WHEN WBC_NVAL_NUM > 19.8 THEN 1 ELSE 0 END AS WBC_GT_198,
            CASE WHEN (TROP_CPK_LAB_LABEL = 'TROPONIN_1'
                        AND TROP_CPK_NVAL_NUM <= 0.04)
                    OR
                        (TROP_CPK_LAB_LABEL = 'CPK_MB'
                        AND TROP_CPK_NVAL_NUM <= 2)
                    THEN 1 ELSE 0
                    END AS TROP_CPK_0,
            CASE WHEN (TROP_CPK_LAB_LABEL = 'TROPONIN_1'
                        AND TROP_CPK_NVAL_NUM BETWEEN 0.05 AND 0.1)
                    OR
                        (TROP_CPK_LAB_LABEL = 'CPK_MB'
                        AND TROP_CPK_NVAL_NUM BETWEEN 3 AND 5)
                    THEN 1 ELSE 0
                    END AS TROP_CPK_2,
            CASE WHEN (TROP_CPK_LAB_LABEL = 'TROPONIN_1'
                        AND TROP_CPK_NVAL_NUM BETWEEN 0.11 AND 0.2)
                    OR
                        (TROP_CPK_LAB_LABEL = 'CPK_MB'
                        AND TROP_CPK_NVAL_NUM BETWEEN 6 AND 10)
                    THEN 1 ELSE 0
                    END AS TROP_CPK_4,
            CASE WHEN (TROP_CPK_LAB_LABEL = 'TROPONIN_1'
                        AND TROP_CPK_NVAL_NUM BETWEEN 0.21 AND 0.3)
                    OR
                        (TROP_CPK_LAB_LABEL = 'CPK_MB'
                        AND TROP_CPK_NVAL_NUM BETWEEN 11 AND 34)
                    THEN 1 ELSE 0
                    END AS TROP_CPK_8,
            CASE WHEN (TROP_CPK_LAB_LABEL = 'TROPONIN_1'
                        AND TROP_CPK_NVAL_NUM >0.3)
                    OR
```

```
                        (TROP_CPK_LAB_LABEL = 'CPK_MB'
                        AND TROP_CPK_NVAL_NUM > 34)
                    THEN 1 ELSE 0
                    END AS TROP_CPK_13
    FROM JGARDNER.TABAK_SCORE TS
    ]';

    --DBMS_OUTPUT.PUT_LINE(sql_string);
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted.');

commit;

END;
```

APPENDIX D – SQL SCRIPTS TO IDENTIFY PREGNANT POPULATION AND ASSOCIATED VARIABLES

```
CREATE OR REPLACE PROCEDURE "JGARDNER"."PREG_FACT_TABLE_PROC_DEID"

AS
--nightherondata.visit_dimension_2 is date shifted
--heronloader.visit_dimension is NOT date shifted
--nightherondata.patient_dimension is NOT date shifted
--heronloader.patient_dimension is NOT date shifted

    my_date date;
    rec_count int;
    i int;
    sql_string varchar2(32767);
    table_name varchar2(100);
    temp_table varchar2(100);
    window_start_date date;
    window_end_date date;

    BEGIN

    DBMS_OUTPUT.put_line('Start of PREG_FACT_TABLE_PROC_DEID');
    window_start_date:= to_date('2015/11/01', 'yyyy/mm/dd');
    window_end_date:= to_date('2016/12/31', 'yyyy/mm/dd');

    DBMS_OUTPUT.put_line('window_start_date = '|| window_start_date);
    DBMS_OUTPUT.put_line('window_end_date = '|| window_end_date);
    --DBMS_OUTPUT.put_line('window_start_date - 280 = '|| (window_start_date -
270));

    table_name:='JGARDNER.PREGNANCY_PROJECT_FACT_TABLE';
    select sysdate into my_date from dual;
    dbms_output.put_line('Creating '||table_name||' table at ' || my_date);
    drop_table(table_name);


    --Create table of PREGNANCY_PROJECT_FACT_TABLE:
    temp_table:='JGARDNER.PREGNANCY_PROJECT_FACT_TABLE';
    drop_table(temp_table);
    sql_string := q'[
CREATE TABLE ]'||temp_table||q'# AS
WITH
    VARIABLES(VARIABLE_CD, TAG) AS

--1. SUPERVISION OF PREGNANCY
        (select concept_cd VARIABLE_CD, 'SUPERVISION OF PREGNANCY' TAG
            from BlueHeronData.ITCP_DIAGNOSIS
            where C_FULLNAME = '\UNMCDIAG\ICD10CM\(Z00-Z99) Fact~rmj9\(Z30-
Z39) Pers~awcc\(Z33) Pregnant state\'
```

```
        UNION
        select concept_cd VARIABLE_CD, 'SUPERVISION OF PREGNANCY' TAG
            from BlueHeronData.ITCP_DIAGNOSIS
            where C_FULLNAME = '\UNMCDIAG\ICD10CM\(Z00-Z99) Fact~rmj9\(Z30-
Z39) Pers~awcc\(Z34) Encounte~f8zh\'

--2. COMPLICATIONS OF DELIVERY
        UNION
        select concept_cd VARIABLE_CD, 'COMPLICATIONS OF DELIVERY' TAG
            from BlueHeronData.ITCP_DIAGNOSIS
            where C_FULLNAME = '\UNMCDIAG\ICD10CM\(O00-O9A) Preg~a433\(O60-
O77) Comp~zbt1\'
        union
        select concept_cd VARIABLE_CD, 'COMPLICATIONS OF DELIVERY' TAG
            from BlueHeronData.ITCP_DIAGNOSIS
            where C_FULLNAME = '\UNMCDIAG\ICD10CM\(O00-O9A) Preg~a433\(O80-
O82) Enco~1n2b\'
        union
        select concept_cd VARIABLE_CD, 'COMPLICATIONS OF DELIVERY' TAG
            from BlueHeronData.ITCP_DIAGNOSIS
            where C_FULLNAME = '\UNMCDIAG\ICD10CM\(O00-O9A) Preg~a433\(O85-
O92) Comp~cvwo\'

--3. DISORDERS DURING PREGNANCY
        union
        select concept_cd VARIABLE_CD, 'DISORDERS DURING PREGNANCY' TAG
            from BlueHeronData.ITCP_DIAGNOSIS
            where C_FULLNAME = '\UNMCDIAG\ICD10CM\(O00-O9A) Preg~a433\(O09-
O09) Supe~n4q2\'
        union
        select concept_cd VARIABLE_CD, 'DISORDERS DURING PREGNANCY' TAG
            from BlueHeronData.ITCP_DIAGNOSIS
            where C_FULLNAME = '\UNMCDIAG\ICD10CM\(O00-O9A) Preg~a433\(O10-
O16) Edem~7xf2\'
        union
        select concept_cd VARIABLE_CD, 'DISORDERS DURING PREGNANCY' TAG
            from BlueHeronData.ITCP_DIAGNOSIS
            where C_FULLNAME = '\UNMCDIAG\ICD10CM\(O00-O9A) Preg~a433\(O20-
O29) Othe~2ok2\'
        union
        select concept_cd VARIABLE_CD, 'DISORDERS DURING PREGNANCY' TAG
            from BlueHeronData.ITCP_DIAGNOSIS
            where C_FULLNAME = '\UNMCDIAG\ICD10CM\(O00-O9A) Preg~a433\(O94-
O9A) Othe~id4e\(O98) Maternal~73qv\'
        union
        select concept_cd VARIABLE_CD, 'DISORDERS DURING PREGNANCY' TAG
            from BlueHeronData.ITCP_DIAGNOSIS
            where C_FULLNAME = '\UNMCDIAG\ICD10CM\(O00-O9A) Preg~a433\(O94-
O9A) Othe~id4e\(O99) Other ma~dyic\'

--4.  'PREGNANCY COMPLICATIONS'
        union
```

```
        select concept_cd VARIABLE_CD, 'PREGNANCY COMPLICATIONS' TAG from
BlueHeronData.ITCP_DIAGNOSIS   where C_FULLNAME = '\UNMCDIAG\ICD10CM\(O00-O9A)
Preg~a433\(O30-O48) Mate~b9yp\(O31) Complica~dv5r\'
        union
        select concept_cd VARIABLE_CD, 'PREGNANCY COMPLICATIONS' TAG from
BlueHeronData.ITCP_DIAGNOSIS   where C_FULLNAME = '\UNMCDIAG\ICD10CM\(O00-O9A)
Preg~a433\(O30-O48) Mate~b9yp\(O32) Maternal~2pdg\'
        union
        select concept_cd VARIABLE_CD, 'PREGNANCY COMPLICATIONS' TAG from
BlueHeronData.ITCP_DIAGNOSIS   where C_FULLNAME = '\UNMCDIAG\ICD10CM\(O00-O9A)
Preg~a433\(O30-O48) Mate~b9yp\(O33) Maternal~17i3\'
        union
        select concept_cd VARIABLE_CD, 'PREGNANCY COMPLICATIONS' TAG from
BlueHeronData.ITCP_DIAGNOSIS   where C_FULLNAME = '\UNMCDIAG\ICD10CM\(O00-O9A)
Preg~a433\(O30-O48) Mate~b9yp\(O34) Maternal~3tra\'
        union
        select concept_cd VARIABLE_CD, 'PREGNANCY COMPLICATIONS' TAG from
BlueHeronData.ITCP_DIAGNOSIS   where C_FULLNAME = '\UNMCDIAG\ICD10CM\(O00-O9A)
Preg~a433\(O30-O48) Mate~b9yp\(O35) Maternal~zs1x\'
        union
        select concept_cd VARIABLE_CD, 'PREGNANCY COMPLICATIONS' TAG from
BlueHeronData.ITCP_DIAGNOSIS   where C_FULLNAME = '\UNMCDIAG\ICD10CM\(O00-O9A)
Preg~a433\(O30-O48) Mate~b9yp\(O36) Maternal~kssm\'
        union
        select concept_cd, 'PREGNANCY COMPLICATIONS' TAG from
BlueHeronData.ITCP_DIAGNOSIS   where C_FULLNAME = '\UNMCDIAG\ICD10CM\(O00-O9A)
Preg~a433\(O30-O48) Mate~b9yp\(O40) Polyhydramnios\'
        union
        select concept_cd VARIABLE_CD, 'PREGNANCY COMPLICATIONS' TAG from
BlueHeronData.ITCP_DIAGNOSIS   where C_FULLNAME = '\UNMCDIAG\ICD10CM\(O00-O9A)
Preg~a433\(O30-O48) Mate~b9yp\(O41) Other di~8hg3\'

--5.  'OBSTETRIC ULTRASOUND'
        union
        select CONCEPT_CD VARIABLE_CD, 'OBSTETRIC ULTRASOUND' TAG from
BlueHeronData.CONCEPT_DIMENSION   where CONCEPT_PATH LIKE
'\UNMC\Procedures\cpt?_codes\70000-79999\76500-76999\76801-76857\76801-
76828\%' ESCAPE '?'

--6.  'ABO AND RH GROUP'
        union
        select concept_cd VARIABLE_CD, 'ABO AND RH GROUP' TAG
            from BlueHeronData.concept_dimension
            where concept_path LIKE '\UNMC\Laboratory Results\Blood and body
fluids\Blood typing and transfusion\882-1\%'
            or concept_cd in ('LOINC:882-1','LOINC:884-7','LOINC:77397-
8','LOINC:972-0',
                    'LOINC:34961-3','LOINC:978-7','LOINC:10331-7','LOINC:1305-
2')

--7.  'HCG IN SERUM EIA'
        union
        select concept_cd VARIABLE_CD, 'HCG IN SERUM EIA' TAG
            from BlueHeronData.concept_dimension
```

```
            where concept_path LIKE '\HH\LP29693-6\LP7786-9\20415-6\%'
            or concept_cd in ('LOINC:20415-6','LOINC:80384-1','LOINC:80385-8')
            group by concept_cd, 'HCG IN SERUM EIA'

--8.  'HCG IN URINE'
        union
        select concept_cd VARIABLE_CD, 'HCG IN URINE' TAG
            from  BlueHeronData.concept_dimension
            where concept_path LIKE '\HH\LP29693-6\LP7786-9\2106-3\%'
            or concept_cd in ('LOINC:2112-1','LOINC:2106-3','LOINC:2113-
9','LOINC:2107-1','LOINC:2114-7',
                    'LOINC:25372-4')
            group by concept_cd, 'HCG IN URINE'

--9.  'HCG IN SERUM_PLASMA'
        union
        select concept_cd VARIABLE_CD, 'HCG IN SERUM_PLASMA' TAG
            from BlueHeronData.concept_dimension
            where concept_path LIKE '\HH\LP29693-6\LP7786-9\2118-8\%'
            or concept_cd in ('LOINC:19080-1','LOINC:19180-9','LOINC:2115-
4','LOINC:2110-5','LOINC:2111-3',
                    'LOINC:2118-8','LOINC:2119-6','LOINC:21198-
7','LOINC:20994-0','LOINC:25373-2',
                    'LOINC:34670-0','LOINC:45194-8','LOINC:55869-
2','LOINC:56497-1')
            group by concept_cd, 'HCG IN SERUM_PLASMA'

--10. 'DISORDER OF PREGNANCY'
        union
        select concept_cd VARIABLE_CD, 'DISORDER OF PREGNANCY' TAG from
BlueHeronData.ITCP_DIAGNOSIS   where C_FULLNAME =
'\UNMCDIAG\SNOMEDCT\404684003\250171008\248982007\118185001\77386006\'

--11. 'PATIENT CURRENTLY PREGNANT'
        union
        select concept_cd VARIABLE_CD, 'PATIENT CURRENTLY PREGNANT' TAG from
BlueHeronData.ITCP_DIAGNOSIS   where C_FULLNAME =
'\UNMCDIAG\SNOMEDCT\404684003\250171008\248982007\118185001\77386006\'

--12. 'MULTIPLE PREGNANCY'
        union
        select concept_cd VARIABLE_CD, 'MULTIPLE PREGNANCY' TAG from
BlueHeronData.ITCP_DIAGNOSIS   where C_FULLNAME =
'\UNMCDIAG\SNOMEDCT\404684003\250171008\248982007\118185001\173300003\10600700
6\106008001\16356006\'

--13. 'VISIT OBSTETRIC CLINIC'
        union
        select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10601054\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10601073\%'
```

```
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10601030\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10601051\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10801042\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10601028\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10604005\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10701043\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10701044\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10601052\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10601053\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10801019\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10801009\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10801018\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10601035\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10701045\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10601072\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10801034\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10601034\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10801017\%'
```

```
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10801015\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\10\10601031\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20101054\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20101073\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20101030\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20104007\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20101051\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20103003\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20101028\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20104005\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20101052\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20101055\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20101053\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20501019\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20501009\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20501018\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20101035\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20101072\%'
```

```
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20501034\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20101034\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20501017\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20501015\%'
        UNION select concept_cd VARIABLE_CD, 'VISIT OBSTETRIC CLINIC' TAG from
BlueHeronData.concept_dimension   where concept_path like
'\i2b2\Encounters\Service Areas\20\20101031\%'

--14. 'UTERUS FUNDAL HEIGHT'
        union
        select concept_cd VARIABLE_CD, 'UTERUS FUNDAL HEIGHT' TAG
            from BlueHeronData.concept_dimension
            where concept_path LIKE '\HH\LP29694-4\LP29717-3\LP7830-5\11881-
0\%'

--15. 'FETAL HEART RATE'
        union
        select concept_cd VARIABLE_CD, 'FETAL HEART RATE' TAG
            from  BlueHeronData.concept_dimension
            where concept_path LIKE '\HH\LP29694-4\LP29711-6\LP7800-8\55283-
6\%'

--16. 'FETAL MOVEMENT'
        union
        select concept_cd VARIABLE_CD, 'FETAL MOVEMENT' TAG
            from BlueHeronData.concept_dimension
            where concept_path LIKE '\HH\LP29694-4\LP29711-6\LP7800-8\57088-
7\%'
    ),

    PROVIDERS(PROVIDER_ID, PROVIDER_NAME, PROVIDER_PATH) AS
    (SELECT PROVIDER_ID, NAME_CHAR, PROVIDER_PATH
        FROM BLUEHERONDATA.PROVIDER_DIMENSION PD
        WHERE REGEXP_LIKE(PD.NAME_CHAR, '*[^ 0123456789 ]')),

    PATIENTS_WITHIN_50_MILES(PATIENT_NUM) AS
    (SELECT DISTINCT PATIENT_NUM
        FROM BLUEHERONDATA.OBSERVATION_FACT
        WHERE CONCEPT_CD IN ('DEM|GEO|UNMC:50mi',
            'DEM|GEO|UNMC:10mi',
            'DEM|GEO|UNMC:15mi',
            'DEM|GEO|UNMC:20mi',
            'DEM|GEO|UNMC:5mi')
        )
```

```
SELECT ENCOUNTER_NUM,
        O.PATIENT_NUM,
        O.CONCEPT_CD,
        --PROVIDER_ID,
        P.PROVIDER_NAME,
        P.PROVIDER_PATH,
        O.START_DATE,
        O.MODIFIER_CD,
        O.INSTANCE_NUM,
        O.VALTYPE_CD,
        O.TVAL_CHAR,
        O.NVAL_NUM,
        O.VALUEFLAG_CD,
        O.QUANTITY_NUM,
        O.UNITS_CD,
        O.END_DATE,
        --LOCATION_CD,
        LD.NAME_CHAR AS ENC_LOCATION,
        O.UPLOAD_ID,
        V.VARIABLE_CD,
        V.TAG,
        P50.PATIENT_NUM AS MILES
        FROM BLUEHERONDATA.OBSERVATION_FACT O
        LEFT JOIN BLUEHERONDATA.LOCATION_DIMENSION LD
            ON O.LOCATION_CD = LD.LOCATION_CD
        LEFT JOIN PROVIDERS P
            ON O.PROVIDER_ID = P.PROVIDER_ID
        JOIN VARIABLES V
            ON O.CONCEPT_CD = V.VARIABLE_CD
        LEFT JOIN PATIENTS_WITHIN_50_MILES P50
            ON O.PATIENT_NUM = P50.PATIENT_NUM
        #';

    --dbms_output.put_line(sql_string);
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);


--PREG_FACT_TABLE_PROC_DEID UPDATE POSITIVE HCG URINE:
    sql_string := q'[
    UPDATE ]'||temp_table||q'[
        SET TAG = 'HCG IN URINE - POS'
        WHERE   TAG = 'HCG IN URINE'
                AND modifier_cd = '@'
                AND valtype_cd = 'T'
                AND tval_char IN
('POSITIVE','P','pos','Pos','POS','pos.','POS.','posative','POSATIVE','posatvi
e','POSISTIVE','positive','Positive','POSITIVE','POSITIVEx2','positivie','posi
tve','POSITVE','positvie','postitive','Postitive','postive','POSTIVE','preg','
+')
        ]';

    --dbms_output.put_line(sql_string);
    execute immediate sql_string;
```

```
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);

--PREG_FACT_TABLE_PROC_DEID UPDATE NEGATIVE HCG URINE:
    sql_string := q'[
    UPDATE ]'||temp_table||q'[
        SET TAG = 'HCG IN URINE - NEG/UNK'
        WHERE   TAG = 'HCG IN URINE'
                AND modifier_cd = '@'
                AND valtype_cd = 'T'
                AND TVAL_CHAR IN
('n','N','nagative','NAGATIVE','neagative','NEAGATIVE','neagtive','Neagtive','
neg','Neg','NEG','neg.','Neg.','NEG.','negaative','NEGAATIVE','negaavie','NEGA
GIVE','negaitve','negaive','Negaive','negarive','negartive','negataive','NEGAT
IAVE','negatie','NEGATIE','Negativ','negative','negaTIVE','nEGATIVE','Negative
','NEGATIVE','-NEGATIVE','negative.','negative`','Negative at
1:20','negativer','negativew','negativr','negatve','negatvie','NEGATVIE','nega
vtive','neggative','NEGGATIVE','negitive','Negitive','NEGITIVE','negitve','neg
native','negstive','negtaive','NEGTAIVE','NEGTAIVEX1','negtive','Negtive','NEG
TIVE','neg x 1','ng','NG','ngative','nrgative','Normal','No Specimen
Received','NO SPECIMEN RECEIVED, SEE SERUM PREG REPORT.','Not Done','See
text','See Text','SEE TEXT','unsure','weak posit','WRONG ORDER','PARTIAL
+','<SEE TEXT>','CANCELED BY HIS','CHANGED','Clear','Cloudy','CONTROL
OK','DELETED','E','Faint pos','false
+','inconclusi','Indeterminate','INDETERMINATE','intermedia','INTERMEDIA','low
positi','Moderate')
        ]';

    --dbms_output.put_line(sql_string);
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);



--PREG_FACT_TABLE_PROC_DEID UPDATE POSITIVE HCG URINE:
    sql_string := q'[
    UPDATE ]'||temp_table||q'[
        SET TAG = 'HCG IN URINE - POS'
        WHERE   TAG = 'HCG IN URINE'
                AND modifier_cd = '@'
                AND valtype_cd = 'N'
                AND NVAL_NUM > 5
        ]';

    --dbms_output.put_line(sql_string);
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);

--PREG_FACT_TABLE_PROC_DEID UPDATE NEGATIVE HCG URINE:
    sql_string := q'[
    UPDATE ]'||temp_table||q'[
        SET TAG = 'HCG IN URINE - NEG/UNK'
        WHERE   TAG = 'HCG IN URINE'
                AND modifier_cd = '@'
                AND valtype_cd = 'N'
```

```
                AND NVAL_NUM <= 5
        ]';

    --dbms_output.put_line(sql_string);
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);


--PREG_FACT_TABLE_PROC_DEID UPDATE HCG IN SERUM_PLASMA - POSTIVE:
    sql_string := q'[
    UPDATE ]'||temp_table||q'[
        SET TAG = 'HCG IN SERUM/PLASMA -POS'
        WHERE   TAG = 'HCG IN SERUM_PLASMA'
                AND modifier_cd = '@'
                AND valtype_cd = 'T'
                AND tval_char IN ('pos','Positive','POSITIVE')
        ]';

    --dbms_output.put_line(sql_string);
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);


--PREG_FACT_TABLE_PROC_DEID UPDATE HCG IN SERUM_PLASMA - NEGATIVE/UNKNOWN:
    sql_string := q'[
    UPDATE ]'||temp_table||q'[
        SET TAG = 'HCG IN SERUM/PLASMA -NEG/U'
        WHERE   TAG = 'HCG IN SERUM_PLASMA'
                AND modifier_cd = '@'
                AND valtype_cd = 'T'
                AND TVAL_CHAR IN ('neg','negative','Negative','<SEE
TEXT>','E',
                    'Equicical, suggest repeating.','Incorrect
Order','PENDING',
                    'See note','See Text','SEE TEXT', 'NEGATIVE')
        ]';

    --dbms_output.put_line(sql_string);
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);


--PREG_FACT_TABLE_PROC_DEID UPDATE HCG IN SERUM_PLASMA - POSTIVE:
    sql_string := q'[
    UPDATE ]'||temp_table||q'[
        SET TAG = 'HCG IN SERUM/PLASMA -POS'
        WHERE   TAG = 'HCG IN SERUM_PLASMA'
                AND modifier_cd = '@'
                AND valtype_cd = 'N'
                AND NVAL_NUM > 5
        ]';

    --dbms_output.put_line(sql_string);
    execute immediate sql_string;
```

```
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);

--PREG_FACT_TABLE_PROC_DEID UPDATE HCG IN SERUM_PLASMA - NEGATIVE/UNKNOWN:
    sql_string := q'[
    UPDATE ]'||temp_table||q'[
        SET TAG = 'HCG IN SERUM/PLASMA -NEG/U'
        WHERE   TAG = 'HCG IN SERUM_PLASMA'
                AND modifier_cd = '@'
                AND valtype_cd = 'N'
                AND NVAL_NUM <= 5
        ]';

    --dbms_output.put_line(sql_string);
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);



--PREG_FACT_TABLE_PROC_DEID UPDATE FETAL HEART RATE - POSITIVE:
    sql_string := q'[
    UPDATE ]'||temp_table||q'#
        SET TAG = 'FETAL HEART RATE - POS'
        WHERE   TAG = 'FETAL HEART RATE'
                AND modifier_cd = '@'
                AND valtype_cd = 'T'
                    AND (upper(TVAL_CHAR) IN ('PRESENT', 'POS', 'PRESENT ',
'PRSENT', 'PRES', 'PRESENT/PRESEMT','PRESE',
                            'PRESETN', 'PRESEMT', 'PRESENT ON US', 'PRESNT',
'PRESENT ON ULTRASOUND', 'PRESNET',
                            'PREENT', 'PRESEENT', 'PRESESNT', 'PRESENET',
'PRESENTS', 'PRESET')
                    OR TVAL_CHAR LIKE '%+%'
                    OR UPPER(TVAL_CHAR) LIKE '%POS%'
                    OR REGEXP_LIKE(TVAL_CHAR, '*[1-9]'))

        #';
    --dbms_output.put_line(sql_string);
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);


--PREG_FACT_TABLE_PROC_DEID UPDATE FETAL HEART RATE - NEGATIVE:
    sql_string := q'[
    UPDATE ]'||temp_table||q'#
        SET TAG = 'FETAL HEART RATE - N/U'
        WHERE   TAG = 'FETAL HEART RATE'
                AND modifier_cd = '@'
                AND valtype_cd = 'T'
                    AND upper(TVAL_CHAR) IN ('ABSENT', 'NEG', 'NOT
HEARD','NONE', 'NEGATIVE', 'NO', 'NOT SEEN',
                            'NOT PRESENT')

        #';
    --dbms_output.put_line(sql_string);
```

```
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);


--PREG_FACT_TABLE_PROC_DEID UPDATE FETAL HEART RATE - POSITIVE:
    sql_string := q'[
    UPDATE ]'||temp_table||q'[
        SET TAG = 'FETAL HEART RATE - POS'
        WHERE   TAG = 'FETAL HEART RATE'
                AND modifier_cd = '@'
                AND VALTYPE_CD = 'N'
                AND NVAL_NUM > 0
        ]';
    --dbms_output.put_line(sql_string);
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);


--PREG_FACT_TABLE_PROC_DEID UPDATE FETAL HEART RATE - NEGATIVE/UNKNOWN:
    sql_string := q'[
    UPDATE ]'||temp_table||q'[
        SET TAG = 'FETAL HEART RATE - N/U'
        WHERE   TAG = 'FETAL HEART RATE'
                AND modifier_cd = '@'
                AND VALTYPE_CD = 'N'
                AND NVAL_NUM = 0
        ]';
    --dbms_output.put_line(sql_string);
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);


--PREG_FACT_TABLE_PROC_DEID UPDATE FETAL MOVEMENT - POSITIVE:
    sql_string := q'[
    UPDATE ]'||temp_table||q'[
        SET TAG = 'FETAL MOVEMENT - POS'
        WHERE   TAG = 'FETAL MOVEMENT'
                AND modifier_cd = '@'
                AND VALTYPE_CD = 'T'
                AND TVAL_CHAR IN ('Increased','Present')
        ]';
    --'Absent','Decreased' are tval_char not included in this logic. Tag from
earlier in procedure remains unchanged.
    --dbms_output.put_line(sql_string);
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);


--PREG_FACT_TABLE_PROC_DEID UPDATE HCG IN SERUM EIA  POSITIVE:
    sql_string := q'[
    UPDATE ]'||temp_table||q'[
        SET TAG = 'HCG IN SERUM EIA - POS'
        WHERE   TAG = 'HCG IN SERUM EIA'
                AND modifier_cd = '@'
```

```
                AND VALTYPE_CD = 'N'
                AND NVAL_NUM > 5
        ]';
    --dbms_output.put_line(sql_string);
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);


--PREG_FACT_TABLE_PROC_DEID UPDATE HCG IN SERUM EIA  NEGATIVE:
    sql_string := q'[
    UPDATE ]'||temp_table||q'[
        SET TAG = 'HCG IN SERUM EIA - N/U'
        WHERE   TAG = 'HCG IN SERUM EIA'
                AND modifier_cd = '@'
                AND VALTYPE_CD = 'N'
                AND NVAL_NUM <= 5
        ]';
    --dbms_output.put_line(sql_string);
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);


--Create PREG_HEALTH_MAIN
    temp_table:='JGARDNER.PREG_HEALTH_MAIN';
    drop_table(temp_table);
    sql_string := q'[
    CREATE TABLE ]'||temp_table||q'# AS
    WITH
    PROVIDERS(PROVIDER_ID, PROVIDER_NAME, PROVIDER_PATH) AS
    (SELECT PROVIDER_ID, NAME_CHAR, PROVIDER_PATH
        FROM BLUEHERONDATA.PROVIDER_DIMENSION PD
        WHERE REGEXP_LIKE(PD.NAME_CHAR, '*[^ 0123456789 ]'))

    SELECT O.ENCOUNTER_NUM,
            O.PATIENT_NUM,
            O.CONCEPT_CD,
            P.PROVIDER_NAME,
            P.PROVIDER_PATH,
            VD.START_DATE,
            O.MODIFIER_CD,
            O.INSTANCE_NUM,
            O.VALTYPE_CD,
            O.TVAL_CHAR,
            O.NVAL_NUM,
            VD.END_DATE,
            LD.NAME_CHAR AS ENC_LOCATION
        FROM BLUEHERONDATA.OBSERVATION_FACT O
        JOIN JGARDNER.PREGNANCY_PROJECT_FACT_TABLE P
            ON O.PATIENT_NUM = P.PATIENT_NUM
        LEFT JOIN BLUEHERONDATA.LOCATION_DIMENSION LD
            ON O.LOCATION_CD = LD.LOCATION_CD
        LEFT JOIN PROVIDERS P
            ON O.PROVIDER_ID = P.PROVIDER_ID
        JOIN BLUEHERONDATA.VISIT_DIMENSION VD
```

```
            ON O.ENCOUNTER_NUM = VD.ENCOUNTER_NUM
        WHERE O.CONCEPT_CD IN
('CPT4:99381','CPT4:99382','CPT4:99383','CPT4:99384','CPT4:99385','CPT4:99386'
,

'CPT4:99387','CPT4:99391','CPT4:99392','CPT4:99393','CPT4:99394','CPT4:99395',
                'CPT4:99396','CPT4:99397')
        AND VD.ENC_TYPE IN ('AV', 'IP', 'EI', 'ED', 'IS')
        GROUP BY O.ENCOUNTER_NUM,
            O.PATIENT_NUM,
            O.CONCEPT_CD,
            P.PROVIDER_NAME,
            P.PROVIDER_PATH,
            VD.START_DATE,
            O.MODIFIER_CD,
            O.INSTANCE_NUM,
            O.VALTYPE_CD,
            O.TVAL_CHAR,
            O.NVAL_NUM,
            VD.END_DATE,
            LD.NAME_CHAR
    #';

-- DBMS_OUTPUT.PUT_LINE(sql_string);
execute immediate sql_string;
DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||table_name||'.');
commit;




/*
drop temporary tables:
    temp_table:='JGARDNER.PREGNANCY_PROJECT_PATIENTS';
        drop_table(temp_table);
    temp_table:='JGARDNER.PREGNANCY_PROJECT_RACE';
        drop_table(temp_table);
    temp_table:='JGARDNER.PREGNANCY_PROJECT_RESIDENCE';
        drop_table(temp_table);
    temp_table:='JGARDNER.PREGNANCY_PROJECT_DEMOGRAPHICS';
        drop_table(temp_table);

*/

commit;

END;
```

```
CREATE OR REPLACE PROCEDURE "JGARDNER"."PREG_FACT_TABLE_PROC"

AS
--nightherondata.visit_dimension_2 is date shifted
```

```
--heronloader.visit_dimension is NOT date shifted
--nightherondata.patient_dimension is NOT date shifted
--heronloader.patient_dimension is NOT date shifted

    my_date date;
    rec_count int;
    i int;
    sql_string varchar2(32767);
    table_name varchar2(100);
    temp_table varchar2(100);
    window_start_date date;
    window_end_date date;

    BEGIN

    DBMS_OUTPUT.put_line('Start of PREG_FACT_TABLE_PROC');
    window_start_date:= to_date('2015/11/01', 'yyyy/mm/dd');
    window_end_date:= to_date('2016/12/31', 'yyyy/mm/dd');

    DBMS_OUTPUT.put_line('window_start_date = '|| window_start_date);
    DBMS_OUTPUT.put_line('window_end_date = '|| window_end_date);
    --DBMS_OUTPUT.put_line('window_start_date - 280 = '|| (window_start_date -
270));

    table_name:='JGARDNER.PREGNANCY_PROJECT_FACT_TABLE';
    select sysdate into my_date from dual;
    dbms_output.put_line('Creating '||table_name||' table at ' || my_date);
    drop_table(table_name);


    --Create table of PREG_FACT_TABLE:
    temp_table:='JGARDNER.PREG_FACT_TABLE';
    drop_table(temp_table);
    sql_string := q'[
    CREATE TABLE ]'||temp_table||q'[ AS
    SELECT ENCOUNTER_NUM,
            PATIENT_NUM,
            CONCEPT_CD,
            PROVIDER_NAME,
            PROVIDER_PATH,
            START_DATE,
            MODIFIER_CD,
            INSTANCE_NUM,
            VALTYPE_CD,
            TVAL_CHAR,
            NVAL_NUM,
            VALUEFLAG_CD,
            QUANTITY_NUM,
            UNITS_CD,
            END_DATE,
            ENC_LOCATION,
            UPLOAD_ID,
            VARIABLE_CD,
            TAG,
```

```
            MILES
        FROM JGARDNER.PREGNANCY_PROJECT_FACT_TABLE@DEID
        ]';

    --dbms_output.put_line(sql_string);
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);


--PREG_FACT_TABLE_DATES:
--Limit facts to date range +/- 90 days
    temp_table:='JGARDNER.PREG_FACT_TABLE_DATES';
    drop_table(temp_table);
    sql_string := q'[
    CREATE TABLE ]'||temp_table||q'[ AS
        SELECT PM.PATIENT_IDE,
                PFT.ENCOUNTER_NUM,
                PFT.PATIENT_NUM,
                PFT.CONCEPT_CD,
                PFT.PROVIDER_NAME,
                PFT.PROVIDER_PATH,
                (PFT.START_DATE - PD.DATE_SHIFT) START_DATE,
                PFT.MODIFIER_CD,
                PFT.INSTANCE_NUM,
                PFT.VALTYPE_CD,
                PFT.TVAL_CHAR,
                PFT.NVAL_NUM,
                PFT.VALUEFLAG_CD,
                PFT.QUANTITY_NUM,
                PFT.UNITS_CD,
                (PFT.END_DATE - PD.DATE_SHIFT) END_DATE,
                PFT.ENC_LOCATION,
                PFT.UPLOAD_ID,
                PFT.VARIABLE_CD,
                PFT.TAG,
                PFT.MILES
        FROM JGARDNER.PREG_FACT_TABLE PFT
        JOIN NIGHTHERONDATA.PATIENT_DIMENSION PD
            ON PFT.PATIENT_NUM = PD.PATIENT_NUM
        JOIN NIGHTHERONDATA.PATIENT_MAPPING PM
            ON PFT.PATIENT_NUM = PM.PATIENT_NUM
        WHERE (PFT.START_DATE - PD.DATE_SHIFT)
            BETWEEN
                (TO_DATE(']' ||window_start_date|| q'[','dd-MON-yy') - 90)
                AND (TO_DATE(']' ||window_end_date|| q'[','dd-MON-yy') + 90)
        ]';

    --dbms_output.put_line(sql_string);
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);


--PREGNANCY_PROJECT_FACT_TABLE_PATIENTS:
    temp_table:='JGARDNER.PREG_FACT_TABLE_PATIENTS';
```

```
    drop_table(temp_table);
    sql_string := q'[
    CREATE TABLE ]'||temp_table||q'[ AS
        SELECT  COALESCE(P.PAT_MRN_ID, P2.PAT_MRN_ID) MRN,
                PPD.PATIENT_IDE ELIGIBLE_PATIENT_IDE,
                PPD.PATIENT_NUM ELIGIBLE_PATIENT_NUM,
                CASE
                    WHEN PPF.PAT_ID IS NOT NULL THEN 'IDENTIFIED'
                    ELSE NULL
                    END AS EPISODE_DATA,
                PPD.AGE,
                PPD.RACE,
                PPD.RESIDENCE,
                PPF.PREG_START_DATE,
                PPF.PREG_END_DATE,
                PPF.PREG_LENGTH,
                PPF.PREG_OUTCOME,
                PPF.GESTATIONAL_AGE,
                PPF.PREG_EPISODE_ID PREGNANCY_EPISODE_ID,
                PPF.DEL_EPISODE_ID DELIVERY_EPISODE_ID,
                EM.ENCOUNTER_IDE,
                PFTD.*
        FROM JGARDNER.PREG_FACT_TABLE_DATES PFTD
        FULL JOIN JGARDNER.PREGNANCY_PROJECT_DEMOGRAPHICS PPD
            ON PFTD.PATIENT_NUM = PPD.PATIENT_NUM
        FULL JOIN JGARDNER.PREGNANCY_PROJECT_FINAL_FILTER PPF
            ON PFTD.PATIENT_IDE = PPF.PAT_ID
        LEFT JOIN NIGHTHERONDATA.ENCOUNTER_MAPPING EM
            ON PFTD.ENCOUNTER_NUM = EM.ENCOUNTER_NUM
        LEFT JOIN PATIENT@VCLARITY P
            ON PFTD.PATIENT_IDE = P.PAT_ID
        LEFT JOIN PATIENT@VCLARITY P2
            ON PPD.PATIENT_IDE = P2.PAT_ID
        ]';

    --dbms_output.put_line(sql_string);
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);

/*Certain elements have been redacted from this procedure as they contain
proprietary content. Further details available via request to author and
demonstration of authorization to view Clarity content.
*/

--PREG_FACT_TABLE_ENROLL:
--Limit to patients seeking consistent care at UNMC
    temp_table:='JGARDNER.PREG_FACT_TABLE_ENROLL';
    drop_table(temp_table);
    sql_string := q'[
    CREATE TABLE ]'||temp_table||q'[ AS
    SELECT ENCOUNTER_NUM,
            PATIENT_NUM,
            CONCEPT_CD,
            PROVIDER_NAME,
```

```
            PROVIDER_PATH,
            START_DATE,
            MODIFIER_CD,
            INSTANCE_NUM,
            VALTYPE_CD,
            TVAL_CHAR,
            NVAL_NUM,
            END_DATE,
            ENC_LOCATION
        FROM JGARDNER.PREG_HEALTH_MAIN@DEID
    ]';

    --dbms_output.put_line(sql_string);
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);

--PREG_FACT_TABLE_ENROLL_DATES:
--Limit to patients seen 2x in 2015-2016 OR have health maintenance visit in
this range
    temp_table:='JGARDNER.PREG_FACT_TABLE_ENROLL_DATES';
    drop_table(temp_table);
    sql_string := q'[
    CREATE TABLE ]'||temp_table||q'[ AS
        SELECT PM.PATIENT_IDE,
                PFT.ENCOUNTER_NUM,
                PFT.PATIENT_NUM,
                PFT.CONCEPT_CD,
                PFT.PROVIDER_NAME,
                PFT.PROVIDER_PATH,
                (PFT.START_DATE - PD.DATE_SHIFT) START_DATE,
                PFT.MODIFIER_CD,
                PFT.INSTANCE_NUM,
                PFT.VALTYPE_CD,
                PFT.TVAL_CHAR,
                PFT.NVAL_NUM,
                (PFT.END_DATE - PD.DATE_SHIFT) END_DATE,
                PFT.ENC_LOCATION
        FROM JGARDNER.PREG_FACT_TABLE_ENROLL PFT
        JOIN NIGHTHERONDATA.PATIENT_DIMENSION PD
            ON PFT.PATIENT_NUM = PD.PATIENT_NUM
        JOIN NIGHTHERONDATA.PATIENT_MAPPING PM
            ON PFT.PATIENT_NUM = PM.PATIENT_NUM
        WHERE (PFT.START_DATE - PD.DATE_SHIFT)
            BETWEEN
                TO_DATE('2015/01/01', 'yyyy/mm/dd')
                AND TO_DATE(']' ||window_end_date|| q'[','dd-MON-yy')
        ]';

    --dbms_output.put_line(sql_string);
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||temp_table);


--Create PREG_ENROLLED_PATIENT_TABLE
```

```
    temp_table:='JGARDNER.PREG_ENROLLED_PATIENT_TABLE';
    drop_table(temp_table);
    sql_string := q'[
    CREATE TABLE ]'||temp_table||q'[ AS

WITH
  PATIENTS_SEEN_2X_IN_WINDOW(PATIENT_NUM, FIRST_ADMIT_DATE, LAST_ADMIT_DATE)
AS
    (SELECT
       VD.PATIENT_NUM,
       MIN(VD.ADMIT_DATE) AS FIRST_ADMIT_DATE,
       MAX(VD.ADMIT_DATE) AS LAST_ADMIT_DATE
     FROM
       NIGHTHERONDATA.VISIT_DIMENSION_2 VD
       JOIN NIGHTHERONDATA.ENCOUNTER_MAPPING EM
         ON VD.ENCOUNTER_NUM = EM.ENCOUNTER_NUM
       JOIN NIGHTHERONDATA.PATIENT_MAPPING PM
         ON VD.PATIENT_NUM = PM.PATIENT_NUM
       JOIN NIGHTHERONDATA.PATIENT_DIMENSION PD
         ON VD.PATIENT_NUM = PD.PATIENT_NUM
       JOIN JGARDNER.PREG_FACT_TABLE_PATIENTS_D PFTD
         ON PM.PATIENT_IDE = PFTD.ELIGIBLE_PATIENT_IDE
     WHERE
       VD.ADMIT_DATE BETWEEN
            TO_DATE('2015/01/01', 'yyyy/mm/dd')
            AND TO_DATE('2016/12/31', 'yyyy/mm/dd')
       AND VD.ENC_TYPE IN ('AV', 'IP', 'EI', 'ED', 'IS')
     GROUP BY
       VD.PATIENT_NUM
     HAVING
       MAX(VD.ADMIT_DATE) - MIN(VD.ADMIT_DATE) > 30),

  PATIENTS_HEALTH_MAINTENANCE(PATIENT_NUM, FIRST_ADMIT_DATE, LAST_ADMIT_DATE)
AS
    (SELECT
       P.PATIENT_NUM,
       CASE WHEN MAX(P.START_DATE) - MIN(P.START_DATE) < 365
            THEN MAX(P.START_DATE) -365
            ELSE MIN(P.START_DATE) END
         AS FIRST_ADMIT_DATE,
       MAX(P.START_DATE) AS LAST_ADMIT_DATE
     FROM
       JGARDNER.PREG_FACT_TABLE_ENROLL_DATES P
       JOIN NIGHTHERONDATA.VISIT_DIMENSION_2 VD
        ON P.PATIENT_NUM = VD.PATIENT_NUM
     WHERE
       P.START_DATE BETWEEN
            TO_DATE('2015/01/01', 'yyyy/mm/dd')
            AND TO_DATE('2016/12/31', 'yyyy/mm/dd')
       AND VD.ENC_TYPE IN ('AV', 'IP', 'EI', 'ED', 'IS')
       AND P.CONCEPT_CD IN
('CPT4:99381','CPT4:99382','CPT4:99383','CPT4:99384','CPT4:99385','CPT4:99386'
,
```

```
'CPT4:99387','CPT4:99391','CPT4:99392','CPT4:99393','CPT4:99394','CPT4:99395',
           'CPT4:99396','CPT4:99397')
     GROUP BY
       P.PATIENT_NUM),

  ALL_ENROLLED_PATIENTS(PATIENT_NUM, FIRST_ADMIT_DATE, LAST_ADMIT_DATE) AS
    (SELECT
       COALESCE(A.PATIENT_NUM, B.PATIENT_NUM),
       CASE WHEN A.FIRST_ADMIT_DATE IS NOT NULL
              AND B.FIRST_ADMIT_DATE IS NOT NULL THEN
              LEAST(A.FIRST_ADMIT_DATE, B.FIRST_ADMIT_DATE)
            ELSE COALESCE(A.FIRST_ADMIT_DATE, B.FIRST_ADMIT_DATE)
            END
          AS FIRST_ADMIT_DATE,
        CASE WHEN A.LAST_ADMIT_DATE IS NOT NULL
               AND B.LAST_ADMIT_DATE IS NOT NULL THEN
               GREATEST(A.LAST_ADMIT_DATE, B.LAST_ADMIT_DATE)
             ELSE COALESCE(A.LAST_ADMIT_DATE, B.LAST_ADMIT_DATE)
             END
          AS LAST_ADMIT_DATE
      FROM
        PATIENTS_SEEN_2X_IN_WINDOW A
        FULL OUTER JOIN PATIENTS_HEALTH_MAINTENANCE B
          ON A.PATIENT_NUM = B.PATIENT_NUM)

    SELECT
      PATIENT_NUM,
      FIRST_ADMIT_DATE AS ENR_START_DATE,
      LAST_ADMIT_DATE AS ENR_END_DATE
    FROM
      ALL_ENROLLED_PATIENTS

]';

-- DBMS_OUTPUT.PUT_LINE(sql_string);
execute immediate sql_string;
DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into '||table_name||'.');
commit;


--Create PREG_FACT_TABLE_PATIENTS_E
    temp_table:='JGARDNER.PREG_FACT_TABLE_PATIENTS_E';
    drop_table(temp_table);
    sql_string := q'[
    CREATE TABLE ]'||temp_table||q'[ AS
    SELECT A.*, B.PATIENT_NUM PATIENT_NUM_ENROLL, B.ENR_START_DATE,
B.ENR_END_DATE
    FROM JGARDNER.PREG_FACT_TABLE_PATIENTS_D A
    LEFT JOIN JGARDNER.PREG_ENROLLED_PATIENT_TABLE B
        ON A.PATIENT_NUM = B.PATIENT_NUM
    ]';

    -- DBMS_OUTPUT.PUT_LINE(sql_string);
```

```
    execute immediate sql_string;
    DBMS_OUTPUT.PUT_LINE(SQL%ROWCOUNT || ' rows inserted into
'||table_name||'.');
    commit;
/*

drop temporary tables:
    temp_table:='JGARDNER.PREG_PATIENTS';
        drop_table(temp_table);
    temp_table:='JGARDNER.PREGNANCY_PROJECT_RACE';
        drop_table(temp_table);
    temp_table:='JGARDNER.PREGNANCY_PROJECT_RESIDENCE';
        drop_table(temp_table);
    temp_table:='JGARDNER.PREGNANCY_PROJECT_DEMOGRAPHICS';
        drop_table(temp_table);

*/

commit;

END;
```