Vantage: Exploring Variability in Inpatient Care Through Physicians' Orders

by

Matthew Lenert

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biomedical Informatics

August 10, 2018

Nashville, Tennessee

Approved:

Colin G. Walsh, MD MA

Yevgeniy Vorobeychik, PhD

Randolph A. Miller, MD

ACKNOWLEDGEMENTS

This project would not have been possible without the guidance of my committee: Colin Walsh MD MA, Randolph Miller MD, and Eugene Vorobeychik PhD. All of them have my gratitude for their mentorship. I would like to acknowledge and thank John Wolfe and the Vanderbilt University Medical Center Finance team for providing encounter-specific-inpatient cost figures. The National Library of Medicine (NLM) deserves thanks and acknowledgement for funding my training. I am extremely grateful to the taxpayers who deemed the NLM's training program worthy of funding. I hope to repay this debt in my future work. I would also like to thank my wife, Jennifer, for her exemplary patience, understanding, and support of my pursuit of an advanced degree.

TABLE OF CONENTS

Page

Chapter

LIST OF TABLES

Table Page

LIST OF FIGURES

Chapter I

INTRODUCTION & BACKGROUND

Introduction

The inpatient care two seemingly identical patients receive for the same initial condition can vary significantly within and across hospitals [1–4]. Such variation affects patient outcomes and often increases the cost of care[5–7]. The authors define care variability as the deviation of clinical practice from the best evidenced-based targeted local approaches. Care variability is a nuanced concept, because the present state of the clinical literature may be unclear; and many factors outside of disease alone affect care decisions[8–11]. For example, disease severity, comorbid conditions, safety records and quality of the available providers, social factors, treatment accessibility, and patient values all play a role in determining the course of care[12–16]. Healthcare administrators assess care variability through the laborious process of chart reviews by expert panels of physicians and/or specialists[17]. In a landscape of evolving evidence bases and differing patient preferences, identifying care variability at any scale becomes challenging, especially given the current complexities of diagnosis and treatment.

The ultimate goal of identifying variable care is to improve patient outcomes, while concurrently reducing or maintaining costs. Measuring care variability can help identify and prioritize services for interventions[18]. The intervention of choice for care variability is care standardization[5,14]. Because healthcare systems are complex, different administrators implement care standardization in varying forms and at different scales[4,19]. Local institutions attempt to standardize care practices through clinical pathways, policies, decision support tools, and easily invoked prescriptive sets of physician orders[9,20–22]. Determining what the intervention will be, on which service to intervene, and how to implement the intervention pose formidable challenges to reducing care variability[8,23,24].

Having observed that physicians' orders comprise the primary driver of inpatient care, the authors hypothesized that analyzing those orders could provide unique insights into variability in patient outcomes[25]. The authors tested this hypothesis through a robust (Huber-White variance corrected) regression analysis using features derived from inpatient physician orders. The authors evaluated the performance of their order-based measure against the current standard for measuring care variability (i.e. variation in costs). The following chapters detail the authors' methods and results. The second chapter describes a novel implementation of multiple imputation methods for completing datasets with missing confounders. In the third chapter, the authors explain how they transformed order data into a surrogate measure for care variability. The fourth chapter brings the aforementioned methods together to create a potential new approach for prioritizing hospital services regarding care standardization and/or guideline implementation. Such a prioritization method could allow hospital administrators to monitor all service lines at a health system scale, using order features as an effective surrogate for deviation from expected outcomes[26].

Description and History of Care Variability

The Institute of Medicine (IOM) identified unequal knowledge dissemination as a primary cause of care variability[14]. To this end, the IOM champions the development of order sets, clinical guidelines, and care pathways to ensure that care does "not vary illogically from clinician to clinician or from place to place[14]." Clinical guidelines represent collaborative efforts by groups of expert clinical specialists, economists, and healthcare administrators to establish standards for the diagnosis, prevention, and management of disease[27,28]. Since guidelines can be incomplete, logically contradictory, or nonspecific, the interpretation and implementation of clinical guidelines carry their own set of challenges[17,29]. For example, the Choosing Wisely Initiative attempts to reduce the utilization of unnecessary or ineffective diagnostic tests and therapies[30]. This initiative is encompasses hundreds of nuanced guidelines spanning many different conditions and diseases. The myriad of guidelines, each with an openness to interpretation, make the initiative's implementation and evaluation onerous[31].

Administrators, researchers, and guideline developers require simplified strategies for targeting similar patients, such as grouping by primary diagnosis[8,32]. Such simplified models are imperfect and, paradoxically, often involve complexities. One tool useful for standardization has been the International Classification of Diseases, version nine (ICD-9); it has a hierarchy comprised of approximately 13,000 diagnoses. The current ICD version, ICD-10, has over 68,000 diagnosis codes[33]. Diagnoses in the ICD are unequally defined in terms of specificity[34]. ICD diagnoses may overlap in definition. For example, the diagnostic criteria of Bipolar Disorder and Schizophrenia each have significant overlap. Thus, clinical interpretation can play a big role distinguishing one from the other. The primary diagnosis is also subject to economic considerations. Billing processes and local workflows may prioritize certain diagnoses over others[35–37]. The presence of multiple (comorbid) diagnoses in a patient can further complicate patient grouping. For example, the patient with hypertension and type-1 diabetes will be treated differently than the patient with hypertension alone[38–40]. Thus, differences among patients grouped by diagnosis can significantly confound the assessment of deviation from the optimal spectrum of treatment. The variety of ICD codes often is difficult to work with, motivating larger patient groupings.

Administrators for the Center for Medicare and Medicaid Services (CMS) and other insurers implemented a higher-level schema for grouping patients' disorders—Diagnosis Related Groups (DRG). Each DRG includes a collection of diagnosis codes and procedures that compress tens of thousands of ICD-10 codes into approximately 750 DRGs. Yale's health policy department began developing DRGs as early as 1967[41]. The DRGs formed the backbone of the Prospective Payment System (PPS) enacted in 1983. The PPS paid hospitals a set per patient rate for each DRG regardless of the hospitals actual expenditures caring for the patient. Implementation of the PPS helped control the explosion of healthcare costs that immediately followed the establishment of Medicare in 1965[41,42]. The PPS incentivized hospitals to deliver efficient services that minimized costs[43]. The significant role DRGs play in reimbursement made them a popular method for grouping patients when analyzing variability in care[12,44,45]. DRGs also became the industry standard for measuring patient case-mix[41].

The literature on care variability documents two prevailing surrogate measures of care variability: nursing intensity variability and cost variability[44–48]. Nursing intensity variability metrics capture quality, quantity, and intensity of nursing care rendered to patients billed for the same DRG[47]. Assessing nursing intensity typically requires trained observers, making it difficult to measure through automatic means[49]. That requirement led to the creation of Nursing Intensity Weights (NIW). The NIW are ordinal scores created by an expert panel of nurses that symbolize the relative nursing intensity of the average patient billed under a given DRG[50]. Administrators use NIW for staffing and reporting purposes. The NIW are ill-suited for measuring care variability because they are not unique for different patients billed under the same DRG. Nursing intensity is highly influenced by severity of illness, which may not be captured by NIW[48]. Regular quantitative measurement of nursing intensity on a patient by patient basis is not feasible for busy care providers[51]. Nursing intensity remains impractical to employ as a care variability surrogate.

An alternative for care variability compares the costs of patient admissions coded with the same DRG. Patients with differences in treatments will presumably have differing costs[52]. Similar to nursing intensity, costs are sensitive to severity of illness[44,45]. This means that it may be difficult to know how much variation in cost to expect across a particular set of patients. Because individual line items are not directly billed in the inpatient DRG-based payment setting, costs are currently difficult to measure on a per-patient basis. Many hospitals proficiently track which supplies, medications, and services individual patients receive during their stay. However, nearly all hospitals struggle to attribute employee time costs (clinician or otherwise) to individual patients[51]. The ambiguity surrounding personnel costs reduces the fidelity of cost figures. For hospitals that do not attempt to model costs at the patient level, the amount billed (charges) can provide a reasonable substitute[53]. Almost all hospitals capture charges at the patient level for reimbursement and/or accounting purposes. The authors believe that the ubiquity of charge and/or cost data make cost/charge variability the most practical surrogate for measuring care variability.

The ultimate goal of identifying variable care is to improve patient outcomes, while potentially reducing costs[54]. Therefore, quality metrics should play a role when comparing care variability measures. Variable care is not necessarily care that varies from the average, but care that deviates from the best choices for the individual patient. Length of stay is a common, albeit flawed, measure of healthcare quality that is also related to cost[43]. Medicare defines the length of stay as the number of midnights a patient spends in the hospital. One DRG may naturally take longer to address than others. Therefore, administrators scale (normalize) patients' lengths of stay by dividing them by the average expected lengths of stay related to the patients' primary DRGs[55]. This ratio puts all admissions on the same scale, facilitating comparisons among DRGs. However, length of stay is a noisy metric[56]. Severity of illness and co-morbidities have strong effects on the length of stay[57,58]. External factors, such as staffing and teaching hospital status, can also affect the length of stay at a hospital[59,60]. Different hospitals are subject to different cost and capacity pressures that add variability to the average length of stay between hospitals[43]. Workflow and the utilization of evidence-based care pathways affect the length of stay. Those practices vary between institutions[61]. The variability of lengths of stay both within and between institutions make it an imperfect outcome measure, but it is an outcome that is easily and objectively measurable across all DRGs.

Types of Missing Data and Multiple Imputation

Since length of stay has some highly influential confounders, models predicting length of stay require full datasets (i.e., without missing observations) regarding confounding variables. A confounder with missing observations can bias results and lead to false conclusions in explanatory models[62]. The authors used an explanatory model to evaluate an order-based care variability surrogate. The key severity of illness covariate had observations with missing data that required some form of imputation.

The distribution of missing data within a variable can have several forms. When all values within a covariate are equally likely to be missing, statisticians describe the data as Missing Completely At Random (MCAR). MCAR usually occurs due to unrelated and random processes, such as lost laboratory samples or equipment malfunctions. Another form of missing data is described as Missing At Random (MAR). MAR data can be corrected for using the other observations that are recorded. MAR data points have no association with the outcome of interest. For example, people without health insurance generally do not get their blood pressure taken, but this fact has nothing to do with what their true blood pressure is. Data that is Not Missing At Random (NMAR) cannot be reliably addressed with observed covariates. NMAR data occurs when the value of an observation determines whether the observation is recorded or missing. For example, blood pressure would be NMAR if most individuals with high blood pressure purposefully avoided seeking medical care in order to avoid costly therapy.

A widely accepted means of dealing with missing data uses multiple imputation. Multiple imputation builds regression models from predictors that do not have data missing to predict values for variables with the missing data[63]. Multiple models are built for each variable with missing data by resampling observations with replacement from predictors that are complete. Multiple imputation can generally fill in missing data when it is MCAR or MAR[63]. Multiple imputation can produce biased results with NMAR data[63]. Under those circumstances, it is less useful. Statisticians improved upon early multiple imputation methods with a technique known as predictive mean matching[64]. Traditionally, multiple imputation uses the fitted models to directly predict what the missing value should be. With predictive mean matching, one uses the fitted models to predict all observations for the dataset, whether the observation has that variable missing or not[65]. Then, one compares the predicted value of each missing data point to the predicted values of all known data points. The matching process determines which known observation is most like the missing one, based upon the predictions. One then imputes the known value for the missing value. The open source statistical programming language R contains several sophisticated packages for multiple imputation with predictive mean matching[66,67].

Inpatient Orders: Definition and Previous Uses

No effective surrogate exists for quantitating care variability[23]. Labor-intensive chart review cannot provide a solution for determining where to focus improvement efforts on a health system-wide scale. Surrogate measures such as nursing intensity variability and cost variability either do not scale well or do not accurately capture deviations from optimal courses of care. Ideally, one should create a better surrogate for care variability based on metrics derived from the care provided. A likely candidate in this regard is physicians' orders.

Inpatient physician orders convert clinicians' plans into actions. They comprise the primary means of delegating and communicating work in the hospital[68]. Physician orders convey work plans asynchronously to other clinicians on a patient's care team, such as nurses. Within the hospital, expensive resources (e.g. a Computerized Tomography (CT) scanner) are often centralized for shared use. A physician-generated order for a CT scan initiates a request for use of the scanner. That request can then be prioritized and queued with other requests to efficiently allocate usage of the CT scanner. Physician orders are used to communicate work to specialized and/or centralized departments, such as the pharmacy or the laboratory. A medication order, for example, communicates to the pharmacy the need to review the patient's care from an allergy and toxicology standpoint, pull the medication from the pharmacy inventory, and produce the correct dosage for a nurse to administer[69]. The wide variety of services provided during the course of hospital care necessitates a large catalog of orders. Electronic ordering systems require flexibility to cover the entire range of potential tasks[70].

Most hospitals in the United States have adopted electronic order entry as part of broader electronic health record systems[71]. Before computerized systems, orders were written by hand or on forms[72]. The order data generated by electronic systems are challenging to work with due to their volume, structure, and variety[73]. The care of a single inpatient can generate hundreds of orders over one encounter. At a health system level, the annual number of orders generated can grow into the tens of millions. The structure of order data varies by order entry system vendor. Free text elements within some of the order data fields exhibit the typical difficulties of working with natural language. Structured elements can also be challenging, as orders cover a large variety of care services, and there may be redundancies[74]. Clinicians do not order each orderable item with equal frequency. Chen et. al. found that structured order catalogs follow a power law distribution for frequency of use[75]. To manage that variety, Chen's research analyses excluded all nursing orders and all other orders with fewer than 256 instances per year[76]. Other researchers who have developed order recommendation systems also constrained order variety by focusing on the most common orders[77,78]. An additional problem in working with order data occurs when two different orders may express the same action. For example, an order for "clean the wound" and an order for "wash surgical site" may describe the same care, but are implemented as two different orders. Conceptually redundant orders can also occur with medications when brand names and generic drug names both comprise permissible orders (e.g. Advil and ibuprofen). Similar problems occur when systems fail to reconcile orders for identical dosages that were described differently (e.g., two 5-mg tablets every 6 hours vs one 10-mg tablet every 6 hours). Duplicative separate orders can occur in clinical specialty departments other than the pharmacy. One approach to reducing duplicative orders with separate names involves mapping orders to higher level concepts; this is not new in informatics. Cimino's "The Med" ontology from the 1980's was one of the first comprehensive attempts applying a conceptual poly-hierarchy to understand and reconcile physician orders[79]. However, few institutions have applied state-of-the-art representation methods for management of physician orders[80,81]. Standardized and widely accepted terminologies, such as the Unified Medical Language System

(UMLS), play a role in disambiguating redundant orders, but have yet to see application past a few academic medical centers[79].

Chapter II

A NOVEL IMPLEMENTATION OF MULTIPLE IMPUTATION IN PYTHON

Overview

The authors conducted a retrospective cohort study examining multiple imputation methods for missing data at a single academic institution—Vanderbilt University Medical Center (VUMC). VUMC is located in Nashville, TN. VUMC's Institutional Review Board approved this study under IRB 151156 Modeling and predicting preventable deviations in healthcare access patterns. VUMC is a level-1 trauma center with a dedicated burn unit, and has 758 licensed beds. VUMC has recently had more than 70,000 emergency room visits per year. Adult (age 18-64) VUMC patients, admitted from July 1st of 2013 through December 31st of 2016, who survived their encounter, and were discharged with a primary service of internal medicine made up the cohort. The date cutoffs were selected due to data availability and to avoid confounding effects from major changes at VUMC. VUMC Finance began computing patient-level inpatient encounter costs at the start of the new fiscal year in 2013. On November of 2017, VUMC switched to a new order entry system.

Python is a programming language popular for client-server and commercial applications. It lacks a widely accepted package for multiple imputation with predictive mean matching. While software tools allow R and Python to interact, that approach complicates application design. As a programming language R does not manage memory well and is primarily intended for single-user data analysis and visualization[82]. The authors developed an imputation package that was designed to work with the data frames format from the Pandas package in Python[83]. The Pandas package interfaces well with the widely used Scikit-Learn machine-learning library[84]. Scikit-Learn has been cited in over 10,000 published applications. While software tools allow R and Python to interact, that approach complicates application design. An imputation pipeline in Python would allow for a more seamless interface between user applications and data modeling programs run on the back-end.

As severity of illness is a key confounder for variability of care measurements, the authors' research project required an imputation tool to fill in missing Emergency-room Patient Severity of Illness (EPSI) scores for its patient cohort. Those patients who were not admitted through the emergency room lack EPSI scores. The study thus needed a real or imputed EPSI score to be present for all observations despite the potential for bias[85]. The authors believe the missing EPSI data was NMAR because those patients electively admitted directly to a hospital bed are generally healthier than those patients admitted via the emergency room. For this reason, the authors examined different imputation strategies to test the sensitivity of potential study results to the imputation method. This led to the authors' development of a multiple imputation package using a modular Python architecture. The package attempted to minimize data processing times, set best practices as defaults, and impute high-quality predictions.

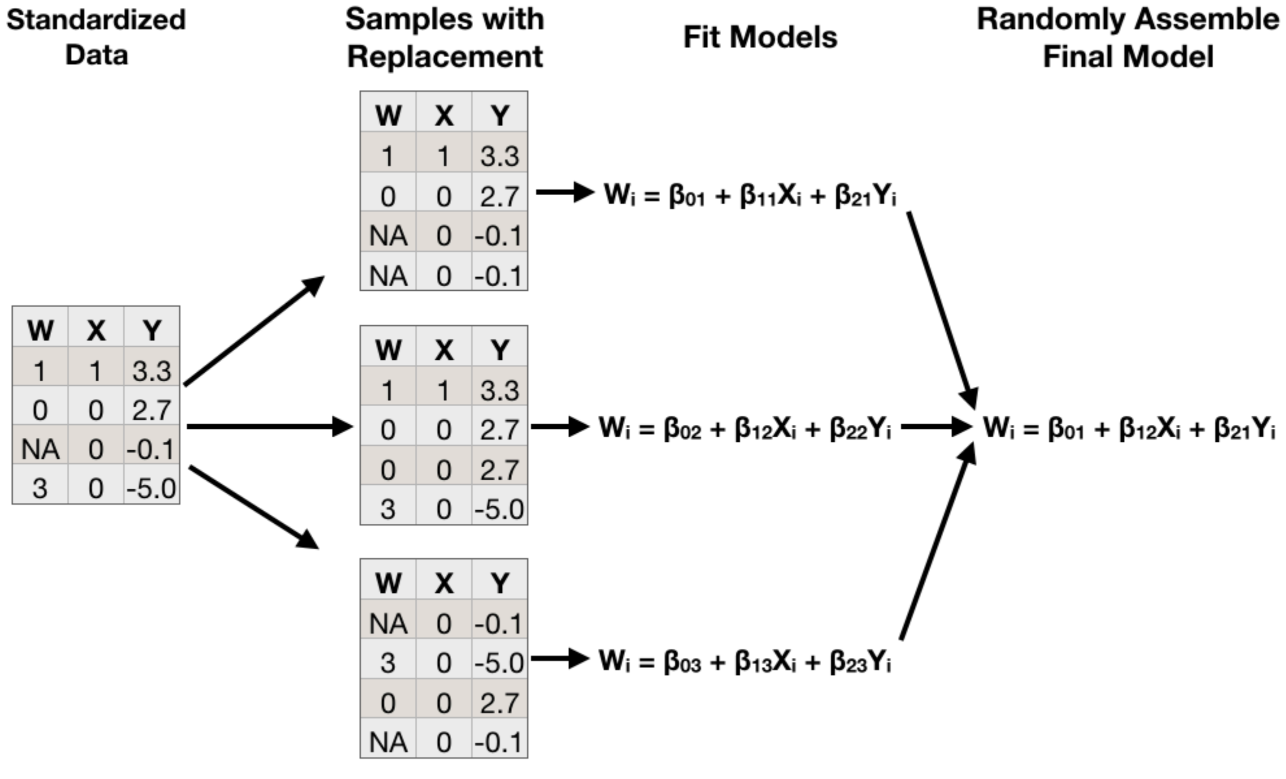
Materials

To develop and evaluate a multiple imputation package, author MCL used a 2015 MacBook Pro with four 2.9 GHz Intel processors and eight GB of RAM and a Linux server with twenty-four Xeon 2.00 GHz cores and 132 GB of RAM. Encounter Data came from VUMC's Research Derivative (RD). The project used Python version 2.7.11. to develop the imputation package, and tested prototypes using the abalone dataset hosted on the University of California Irvine's machine learning data repository[86]. The project also used data from a published suicide-risk model for evaluation[87]. The authors used the Hmisc package in version 3.4.3 of R for comparative purposes.

Methods

The Python imputation package determines which columns have missing data and which do not.

Next, it centers all of the continuous variables' data at zero by subtracting their means. Then it scales the data by dividing by the standard deviation. Penalized regression (used for the imputation) is sensitive to the scale of the data. The Python imputation package splits non-continuous categorical features into binary features, with one category acting as a reference for the others. Next, the Python implementation takes samples with replacement from the data using a feature with missing data as the outcome, and the features with no missing data as its predictors. It fits a LASSO penalized regression for each sample with replacement. The coefficients of the predictors for each model are stored for use later. After fitting all the models, a final model is assembled by randomly sampling the previously saved (posterior distribution of) coefficients for the predictors. This process is illustrated in Figure 1.


Standardized Data
Samples with Replacement
Fit Models
Randomly Assemble Final Model
$W_i = \beta_{01} + \beta_{11}X_i + \beta_{21}Y_i$
$W_i = \beta_{02} + \beta_{12}X_i + \beta_{22}Y_i$
$W_i = \beta_{03} + \beta_{13}X_i + \beta_{23}Y_i$
$W_i = \beta_{01} + \beta_{12}X_i + \beta_{21}Y_i$

W	X	Y
1	1	3.3
0	0	2.7
NA	0	-0.1
3	0	-5.0

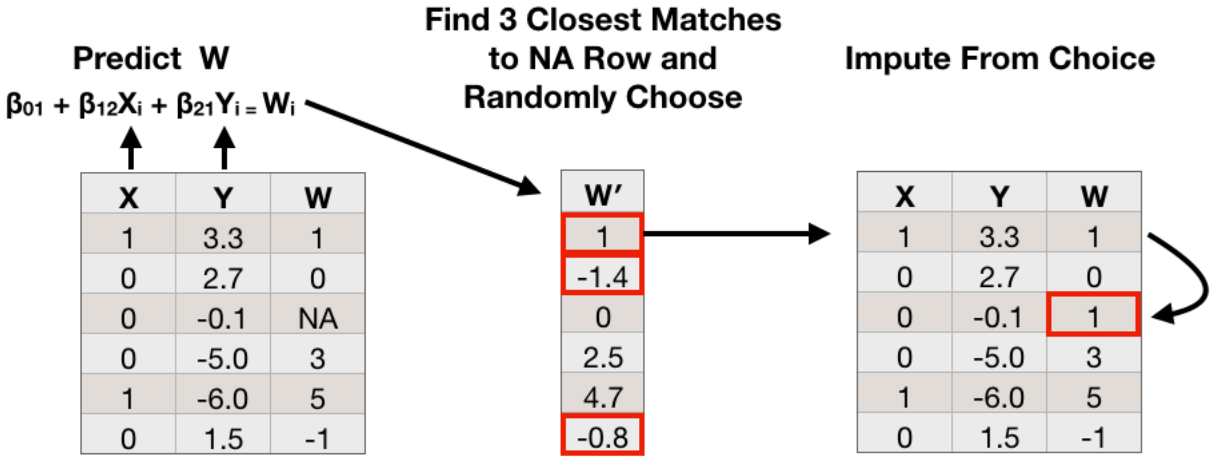
W	X	Y
1	1	3.3
0	0	2.7
NA	0	-0.1
NA	0	-0.1

W	X	Y
1	1	3.3
0	0	2.7
0	0	2.7
3	0	-5.0

W	X	Y
NA	0	-0.1
3	0	-5.0
0	0	2.7
NA	0	-0.1

Figure 1: Multiple Imputation Modeling Process

Predictive mean matching determines the value to be imputed for the missing data. The predictive mean matching algorithm is demonstrated in Figure 2. The entire multiple imputation modeling process with predictive mean matching is repeated for each feature with missing data. The authors' package checks to ensure that there are at least 10 observations per predictor to properly fit the regressions.


Predict W
$\beta_{01} + \beta_{12}X_i + \beta_{21}Y_i = W_i$
Find 3 Closest Matches to NA Row and Randomly Choose
Impute From Choice

X	Y	W
1	3.3	1
0	2.7	0
0	-0.1	NA
0	-5.0	3
1	-6.0	5
0	1.5	-1

W′
1
-1.4
0
2.5
4.7
-0.8

X	Y	W
1	3.3	1
0	2.7	0
0	-0.1	1
0	-5.0	3
1	-6.0	5
0	1.5	-1

Figure 2: Predictive Mean Matching Process

The authors compared their package to an existing “gold standard” multiple imputation method—renowned Professor Frank Harrell’s aregImpute function in the Hmisc R package[66]. Using only default settings for each package, the study compared each package’s imputation results using two datasets. First, the authors used the abalone dataset, which has both continuous and categorical predictors and no missing values[86]. The authors randomly inserted missing values into two of the seven predictors of the dataset. The study repeated this process 100 times to create 100 different test datasets with varying patterns of missing data. Next, the authors’ and gold standard packages imputed values into the 100 test datasets. The authors compared the imputed values to the known values in the full dataset. For the continuous variable the study reported the 95% confidence intervals of the Mean Squared Error (MSE) on the standard deviation scale. The study scaled the MSE, because not all of the continuous variables were on the same scale. For the categorical variables the study reported the 95% confidence intervals for each package’s accuracy.

The second evaluation of the packages used real data from VUMC inpatients with depression. In the depressed patients data set, both functions imputed values for the patients’ Body Mass Index (BMI) based on demographic data, medications, utilization history, and diagnosis codes. The true values of the missing data in that dataset are unknown[87]. The authors examined the distribution of the imputed values in the dataset and compared that distribution to the distribution of non-imputed values. The authors repeated this comparison over different time points (30, 14, 7, and 0 days).

Lastly, the authors further contrasted the two imputation methods by imputing the EPSI scores for the patient cohort of the study. The authors sought to identify differences in the distribution of EPSI scores before and after imputation. This comparison was informal and more of a consistency check with the assumption that directly admitted patients have generally less acute problems than patients admitted through the emergency room.

Results

For the Abalone dataset Harrell's aregImpute recorded accuracy between [46.5%, 47.6%] when imputing for categorical variables. The authors' package imputed the correct category between [29.6%, 31.3%] of the time. The scaled MSE of aregImpute was within [0.0058, 0.0061], while the Python imputer had an MSE between [0.0826, 0.0859]. The aregImpute function consistently bested the authors' package imputing values for the abalone dataset.

In the suicide-risk model patient data set, both functions imputed values for the patients' Body Mass Index (BMI) based on demographic data, medications, utilization history, and diagnosis codes. Figure 3 shows the median and interquartile range of the imputed data points and the un-imputed data.

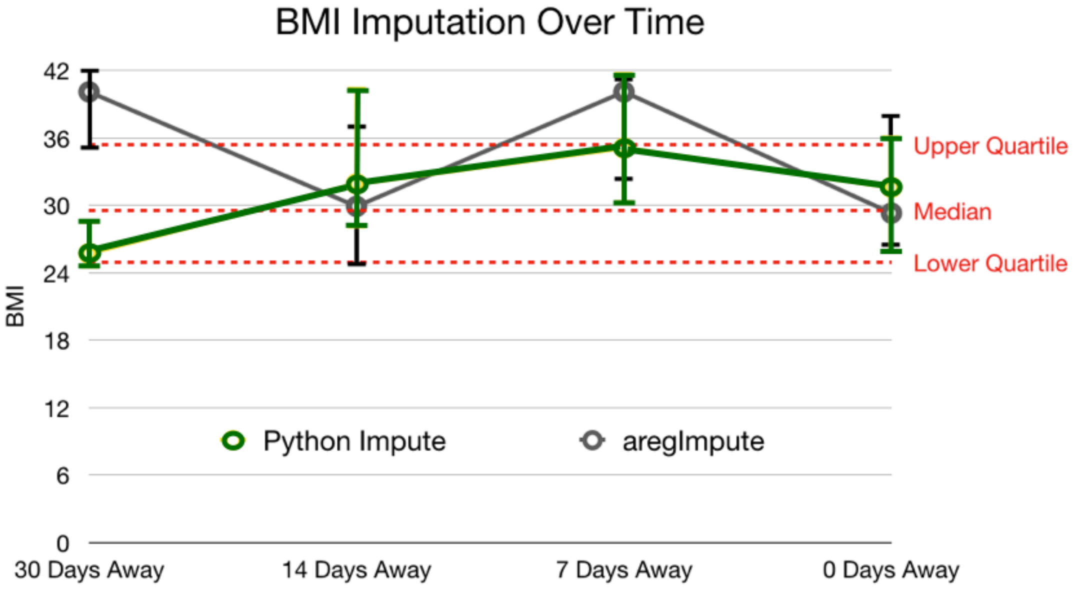


Figure 3: Comparison of Imputed BMI Values

Figure 4 visualizes how each factor of the selection criteria limited the cohort of patient encounters in Vanderbilt's RD. The final cohort was 13,597 inpatient admissions. Of the total cohort 5,303 (39%) encounters did not have an EPSI score.

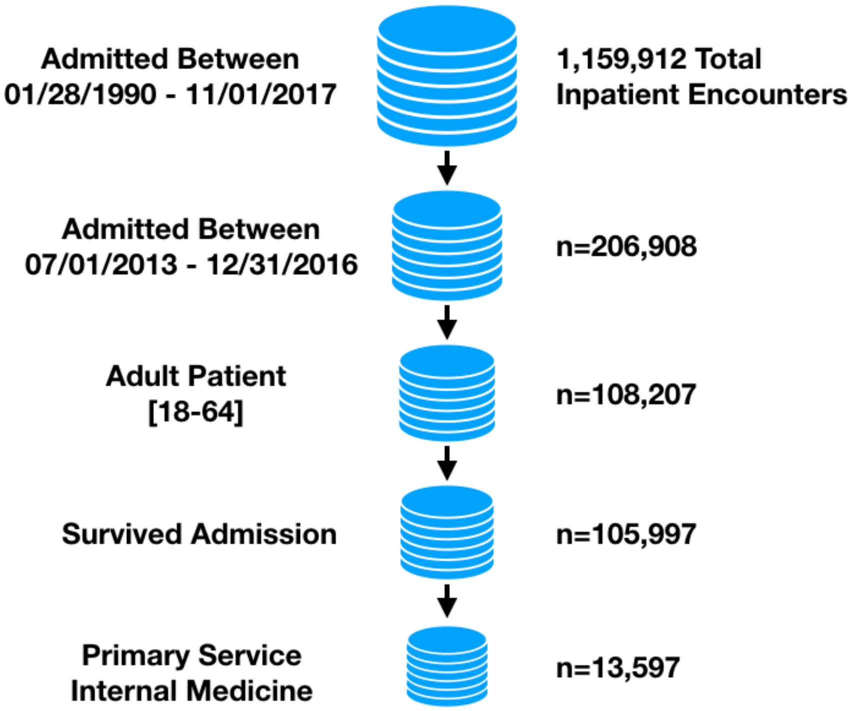


Figure 4: Relationship of Sample Size to Selection Criteria

Both methods imputed EPSI scores for later analysis. Table 1 displays how each imputation method changed the distribution of EPSI scores from the original state.

Imputation Method	Missing	1-Least Severe	2	3	4-Most Severe
aregImpute	0%	13.6%	32.4%	38.4%	15.6%
Python Imputer	0%	7.4%	23.8%	40.3%	28.5%
Neither	39%	7.4%	19.6%	24.7%	9.3%

Table 1: EPSI Distribution of Scores Across Imputation Methods

Discussion & Limitations

The authors' Python Imputer successfully imputed values for two very different datasets without any modification to the programming. This suggests that the Python Imputer should be able to handle the intended use case of EPSI index imputation. The Python Imputer is also one of the first flexible multiple imputation packages for that programming language that the authors are aware of.

The aregImpute package consistently outperformed the Python Imputer for the abalone dataset. The aregImpute package fits cubic splines to the data with a default of three knots. Cubic spline regression provides substantial advantages for relatively simple non-linear relationships that are poorly captured with a linear model. There is no spline regression package in Python, so the authors could not have utilized a similar strategy for their imputer to achieve parity in performance. Building a flexible spline regression function was out of scope for this project due to complexity.

The large number of predictors (over 1,500) in the depression data set posed challenges for imputation. These predictors included counts of prescribed medications by drug class and healthcare utilization, indicators for comorbidities, and demographic information. The aregImpute package would regularly impute BMI values outside the interquartile range of patients with a recorded BMI. The median BMI of the Python Imputer remained within this inter-quartile range over all four-time points. The inter-quartile Python Imputer also had greater overlap with the rest of the data's interquartile range. It is impossible to draw definite conclusions, because the true BMIs are not known. However, the MAR nature of BMI suggested that it should follow patterns similar to that of the rest of the data.

In the original patient cohort dataset, EPSI scores were centered at 3. The Python imputer generated values most often for severity category 4, with the remaining imputations predominantly in severity category 3. The aregImpute package placed the majority of missing observations in the severity categories 2 and severity 3. This more closely matched the expectations of the authors: voluntarily admitted patients more likely would have less severe disease than patients admitted through the emergency room.

The evaluation was limited to three data sets and two imputation packages. Of the three datasets, only one had known values for missing results. The authors used existing datasets instead of creating their own data generation process. The Python Imputer may be inferior to the more complex aregImpute algorithm, as Python lacks a good spline regression package. This evaluation also did not evaluate the usability of either package for comparison. Usability could play a large role in influencing end-user preferences and in determining realized utility.

Chapter III

ENGINEERING FEATURES FROM ORDERS

Overview

The authors constructed a programmatic pipeline (the Vantage system) to process data from VUMC's former (1995-2017) order entry system, WizOrder. Faculty members in the Vanderbilt Department of Biomedical Informatics developed WizOrder as a computerized inpatient order entry system with integrated decision support in 1994-1995 for use at Vanderbilt University Hospital. WizorOrder was rolled out incrementally ward-by-ward with most units adopting the system by 1998[88–91]. After later customization and evolution of the system, Vanderbilt licensed WizOrder to McKesson Corporation for commercial release as Horizon Expert Orders in 2001. The commercial version of the system was "back-installed" at VUMC in 2005-06. At VUMC, clinicians entered patient orders into WizOrder/Horizon Expert Orders (henceforth abbreviated as "WizOrder") until November 1, 2017, when the system was replaced by a different order entry platform. The WizOrder system had 98 structured data elements (numeric, Boolean, dates, or categorical) that comprised each individual order. Structured data elements included fields such as the ordering physician identifier, the service department of the order, and the encounter number. The service department field in WizOrder orders could have sixteen different values: pharmacy, nursing, laboratory, radiology, nutrition, neurology, cardiology, endocrinology, respiratory, pulmonology, vascular, rheumatology, social work, audiology, rehabilitation, and miscellaneous. Users of the WizOrder system could modify, delete, and regenerate orders. Those modifying actions generated additional ancillary orders that referred to the active order that they modified, as shown in Figure 5.

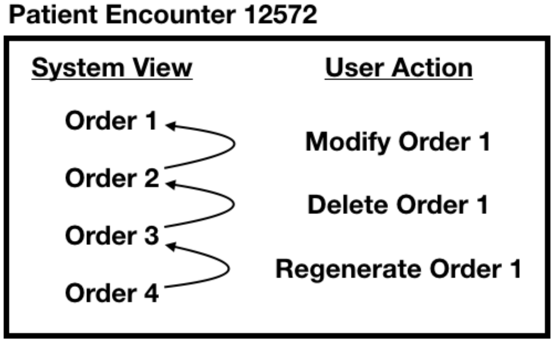


Figure 4: Example of Order Modification, Deletion, and Regeneration

Materials

To extract WizOrer order features, author MCL used a 2015 MacBook Pro with four 2.9 GHz Intel processors and eight gigabytes (GB) of random access memory (RAM) and a Linux server with twenty-four Xeon 2.00 GHz cores and 132 GB of RAM. All data access scripts were written in either Oracle's PL/SQL or IBM's Netezza SQL. All processing scripts were written in Python version 2.7.11. The authors referenced the UMLS Metathesaurus Browser Release 2017AB. Data from WizOrder came from VUMC's Operational Electronic Data Warehouse (EDW). Encounter Data came from VUMC's Research Derivative (RD). Authors obtained Vanderbilt IRB permission to conduct the analyses of orders.

Order Concept Mapping Methods

The authors downloaded all orders for every patient in the previously defined cohort (adult patients, admitted from 7/1/2016-12/31/2016, who survived their encounter, and were discharged with a primary service of internal medicine) from the VUMC EDW. The authors saved the data into a number of large Comma Separated Value files (CSV), due to restrictions in withdrawing large datasets from the data warehouse. WizOrder was occasionally used for data entry for patient parameters (e.g., related to risk of falling, pressure ulcers, etc.) that were not actionable orders per se. The Vantage system eliminated any such nurse-entered patient-specific data items. “Panel orders” packaged a group of individual orders so that they could be generated with a “single” order - e.g., Basic Metabolic Panel, (BMP) included serum sodium, potassium, creatinine, and other items. The comprehensive metabolic panel bundled orders for 13 individual laboratory tests.

The next phase in data processing focused on collapsing functionally similar orders into groupings related to an over-arching UMLS terminology concept. This facilitated the handling of logically redundant orders in a uniform manner. For example, the order to “Elevate HoB” and the order to “Adjust Head of the Bed to 45 degrees” both fall under the UMLS concept “Elevation of head of bed ^C1827414” (The text before the “^” contains the UMLS concept name and the text afterward is the UMLS Concept Unique Identifier, or CUI). The authors used the free-text order comment field for concept mapping to the UMLS (e.g., most nursing orders were of the form “Nursing: [free text comment]”), but broke the problem down by the service department (destination) of the order. This type of breakdown separated the orders by clinical domain area. The authors then applied one of three strategies to map the order descriptions to a terminology concept. The first strategy directly mapped the order description to a terminology concept manually. The second strategy involved parsing the order description for a relevant concept. The third strategy used regular expressions on the order description to find the appropriate concept from a curated list of concepts. Author MCL explored using automated approaches such as MetaMap, but found that its output did not compress order concepts enough[92]. To reach the same results as manual mapping, the authors would have to traverse the UMLS concept hierarchy and explore synonyms for every order description mapped. This iterative process of mapping and tree traversal would have added considerable complexity to the project.

The authors directly mapped order descriptions in clinical domains with a relatively small number of unique order descriptions that were unstructured. For the direct mapping strategy author MCL attempted to compress order descriptions into higher level concepts by manually mapping order descriptions using the UMLS MetaThesaurus Browser to find the best concept. Author MCL first reviewed a list of previously used UMLS concepts before adding any new concepts. In a spreadsheet, author MCL labeled each unique order description (1 per row) with a UMLS concept. This spreadsheet formed the basis for a dictionary that would perform the labeling systematically. The authors attempted to constrain the number of UMLS terminologies used to the minimum needed for coverage. Two board-certified internal medicine physicians (authors CGW and RAM) reviewed random samples of mapped concepts to validate their accuracy. Both reviewers had to approve each mapping for it to be counted as correct. The authors recorded the jointly reviewed accuracy scores (number correct/number reviewed) for the samples reviewed.

The string parsing mapping strategy used string substitution and pattern matching to take highly structured order descriptions and standardize them. The authors used pattern matching to remove unnecessarily specific words in descriptions, such as removing “Bilateral” in an order description of “Computerized Tomography Lung Bilateral” to form “Computerized Tomography Lung.” Radiology orders were reduced to a modality term and a location term. The authors used string substitution to standardize synonyms. For example some laboratory orders had a specimen type of serum, and some a specimen type of blood. These terms are not synonymous but were felt to be sufficiently similar clinically for the purposes of this project, so the authors replaced “blood” with “serum.” Laboratory orders were modified until an analyte and specimen type remained or the name of the laboratory panel remained. The authors used a Python script to parse the structured order descriptions. Again the physician authors reviewed random samples of the concept mappings for accuracy. The study reported the accuracy scores, where both reviewers had to agree.

Domains with large numbers of unique free-text order descriptions necessitated the use of regular expressions for concept mapping. The authors first formed a list of UMLS noun concepts to cover the majority of descriptions. Next, the authors developed a list of UMLS verb/adjective concepts.

The verb concepts modify the noun concepts. The authors then developed a list of regular expressions for each UMLS noun and verb concept listed. The regular expressions mapped to a concept would serve as the linking mechanism from the concept to an order description. The regular expressions were applied in two phases. The first phase searched for matches in an order description to regular expressions tied to a noun UMLS concept. The second phase matched a verb/adjective concept to an order description, if and only if the noun concept had applicable verb concepts. For example, the Vantage system would map an order of “Flush Feeding Tube” to the “Feeding tube^C2945625” concept in phase 1, and then have the “Irrigate or flush system^C0512622” concept appended, resulting in “Feeding tube^C2945625|Irrigate or flush system^C0512622.” The authors sampled order descriptions by concept, to validate the accuracy of all order descriptions matched by the regular expressions. Two physician reviewers evaluated all the order descriptions mapped to the concepts in the random sample. Again both reviewers had to agree on each order description mapping. Concepts with no errors in any of the order descriptions mapped to that concept were labeled accurate. The study reported the concept level and order description level accuracies for each sample. The list of UMLS noun concepts may not cover all order descriptions. The list of regular expressions per concept may not include all relevant order descriptions. Therefore, any order description that was unmatched to a concept had its description used at its terminology concept.

After iteratively correcting and evaluating all of the concept mappings, the authors applied the mappings to all of the order data. The direct mapping and string parsing strategies used dictionaries for mapping. The order description of an order was used as the key in the dictionary, which then returned the mapped UMLS concept. The regular expression method was run over each order description. Next, the authors indexed modifying-action orders (revise, discontinue, regenerate) to their original order. Using the index, action orders inherited the UMLS concepts from the original order. As an example, a “Telemetry” order was eventually followed by a “Discontinue Telemetry” order before discharge. The

"Discontinue Telemetry" order would inherit the "Cardiac monitoring^C0150496" concept from the original "Telemetry" order and have the action modifier "Discontinued^C1444662" appended, resulting in "Cardiac monitoring^C0150496|Discontinued^C1444662."

Order Feature Methods

With most of the orders mapped to terminology concepts, the authors attempted to derive order-based features suitable for statistical modeling. The authors analyzed care variability using statistical models, which is detailed in the next chapter. Most statistical models require data to be numerically formatted in matrix representation. This representation can support predictors that are numeric (continuous or not), categorical, and ordinal. Thus, the authors needed their order-derived features to be numeric and reflective of some aspect of care variability. For example order features can express the number of unique orders, the intensity of care, or the set of orders that make up common practice.

The authors sought to utilize the number of intended executions of an order as a numeric feature. Every order has a categorical frequency and a numeric duration. The authors algorithmically mapped categorical frequencies into standardized rates of: completions/day, completions/week, and completions/month. The authors assumed that "as needed" also known as "prn" orders were always executed at the specified duration and frequency; no clear way existed to know how many times each order was actually carried out since the study dataset lacked nursing medication administration records. Determining actual order durations also required some processing, because users could queue orders for long periods of time into the future, and intervening events could subtly affect actual order duration. Estimation of the true duration required comparison of the order start date to the discharge date, as well as to any modifying (child) orders. Figure 6 illustrates several different scenarios that could arise when calculating an order's duration.

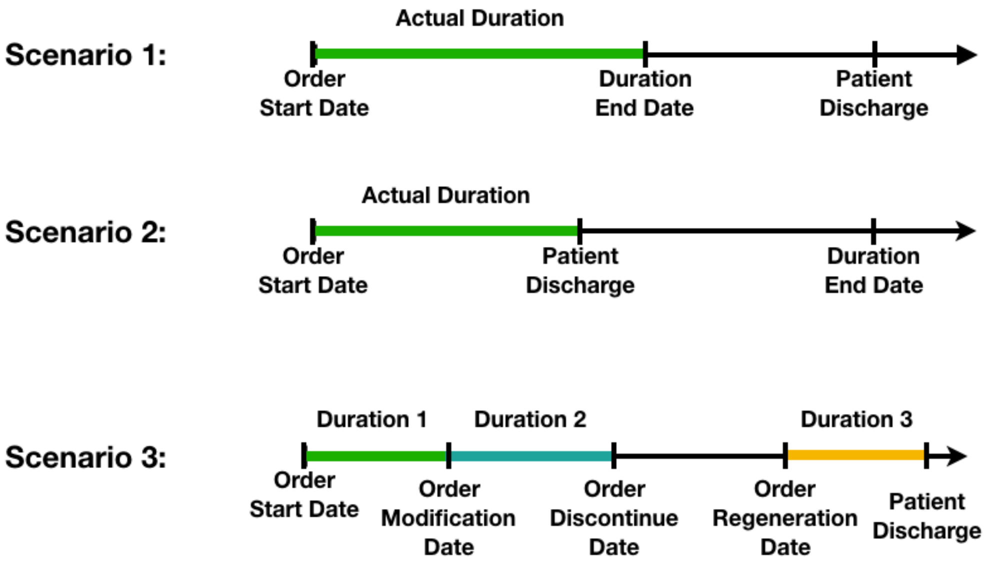


Figure 5: Select Examples of Order Duration Calculation

Order Grouping Methods

The last aspect to deriving potentially predictive features from orders involved grouping. Aggregating order features into higher-level groups is important for analysis of orders across various criteria. The most granular grouping is at the individual order level. Order data can be grouped by terminology concept, service department, or ordering provider ID. Next, order level groupings can be averaged over different time scales. One can either average order groupings by day from admission or by encounter. Grouping by day averaged order features for each day of the admission. Grouping by encounter provides an encounter average for each order feature, which is then scaled by the length of stay. The encounter grouping scaled order features by the length of stay to control for differences in the length of stay between encounters. Patient characteristic groupings occur after time-scale groupings. For example, users can group by patient age at admission, discharge diagnosis, or admitting service. Figure 7 demonstrates different groupings and how those groupings can decrease or increase unique values in the data.

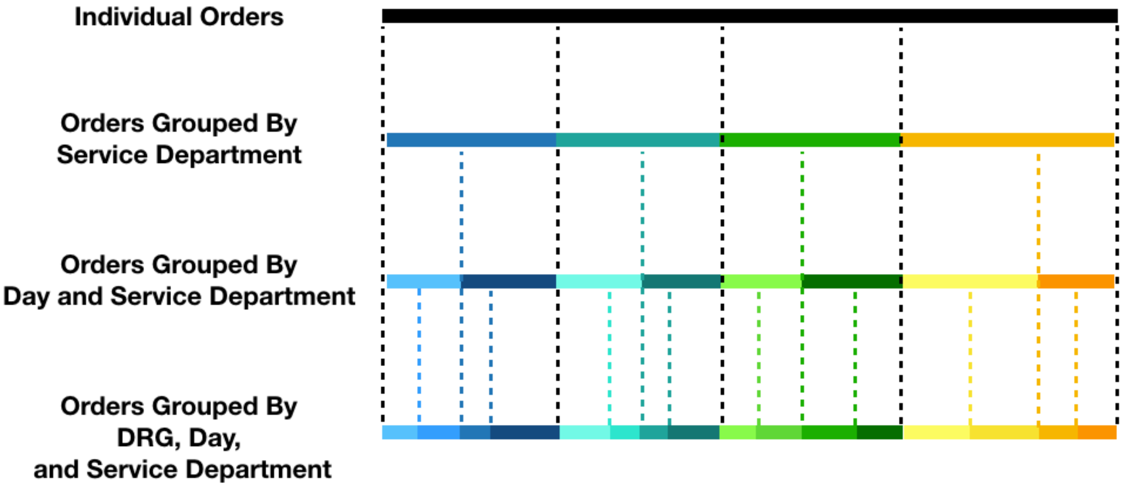

Individual Orders
Orders Grouped By Service Department
Orders Grouped By Day and Service Department
Orders Grouped By DRG, Day, and Service Department

Figure 6: Effect of Multiple Order Groupings on Category Variety

Concept Mapping Results

The authors initially downloaded 21,198,137 orders (11.6 GB) from all the patients included in the cohort. This initial download included orders and nurse-entered data for the individual patients that made up the cohort but it also included their encounters from outside the cohort time window. The Vantage system filtered, processed, extracted, and created features; then, grouped the order data in less than nine hours using ten processors on a Linux server with twenty-four Xeon 2.00 GHz cores. Applying the regular expressions for concept mapping was the rate-limiting step in the pipeline. Each processing step was lossless after the initial filter by encounter number and the elimination of nurse-entered data. After processing, 3,455,292 orders (1.1 GB) remained for the encounters in the cohort. Of the 17,742,845 orders excluded, 17,293,657 orders came from encounters outside of the cohort time window for patients who had at least one encounter inside the cohort time window. Finally, the system excluded 449,188 entries comprised of nurse-entered patient-specific data.

The authors applied a variety of concept mapping strategies, based on each order's clinical domain and other characteristics of the orders. Most of the Pharmacy orders already (before downloading the EDW dataset) had been mapped to the RxNorm terminology by Vanderbilt's Research

Derivative Team using the Medex system[93]. The authors manually mapped 87,408 order descriptions to UMLS concepts. Of the total 87,174 manually mapped order descriptions, only 10,234 were pharmacy orders. The remaining manually mapped orders included all of the unique order descriptions from twelve different service departments. Laboratory and Radiology orders were well structured. This made them the ideal candidates for concept mapping by string parsing. Nursing orders had the greatest variety, making them the best candidate for regular expression-based mapping. Table 2 presents the number of unique order descriptions by service department and the terminologies and strategies applied.

Service Department	Number of Unique Orders	Terminology Applied	Mapping Strategy
Pharmacy	678,820	RxNorm [94]	Direct Mapping
Laboratory	56,663	Custom (Analyte and Specimen Type)	String Parsing
Radiology	8,677	Custom (Location and Modality)	String Parsing
Nursing	485,858	SNOMED-CT [95], MEDCIN[96], HCPCS[97]	Regular Expressions
All Others	77,174	SNOMED-CT [95], MEDCIN[96], HCPCS[97]	Direct Mapping

Table 2: Terminologies and Strategies Applied to Order Service Departments

For the nursing orders, the authors identified 305 unique noun UMLS concepts and 52 unique verb/adjective UMLS concepts. The authors coded 695 regular expressions for the 305 noun concepts and 171 regular expressions for the 52 action concepts. The sampling and evaluation statistics for the order concept mapping done by direct mapping and string parsing are presented in Tables 3.

Service Department	Mapping Strategy	Number of Orders Mapped	Sample Size	Accuracy
Pharmacy	Direct Mapping	10,234	200 (2%)	99%
All Others (Excludes Nursing)	Direct Mapping	77,174	200 (0.3%)	99%
Laboratory	String Parsing	56,663	500 (1%)	84%
Radiology	String Parsing	8,677	500 (6%)	90%

Table 3: Expert Validation of Directly Mapped and String Parsed Orders

The authors sampled the nursing orders by concept. All the order descriptions within the same concept needed to be approved by the physician reviewers for that concept to be labeled as accurate. The validation and sampling statistics for concept mapping done by regular expressions is shown in Table 4.

Sample #	Number of Concepts Sampled	Accurately Mapped Concepts	Number of Orders in Sample	Order Mapping Accuracy
1	36/276 (13%)	23/36 (64%)	2,945/485,858 (1%)	93%
2	42/313 (13%)	28/42 (67%)	3,319/485,858 (1%)	97%
3	20/305 (7%)	11/20 (55%)	481/485,858 (0.1%)	85%

Table 4: Expert Validation of Nursing Orders Mapped by Regular Expressions

Order Feature Results

From the grouped order data the authors derived different potential features to make up an order variability metric. Table 5 enumerates the different order features derived for each DRG-service department group. The authors used different central moments of the distribution (variance (2nd) and kurtosis (4th)) of the intended completions/day to quantify differences between encounters within a DRG/service-department group. The unique order count describes the variety of orders used across all encounters in a DRG/service-department group. The common order count and common order ratio both quantify how many orders are held in common across encounters within a group. To derive the common order count, the authors create a set of unique orders for each encounter. Next, the authors determine which orders are present in more than half of all the encounters coded for the same DRG.

Feature Name	Description
Kurtosis of Completions/Day	Average of the Average Number of Order Completions per Day
Variance of Completions/Day	Variance of the Average Number of Order Completions per Day
Unique Order Count	Count of Unique Terminology Concepts Across All Encounters
Common Order Count	Count of Terminology Concepts Present in > 50% of Encounters for the Group
Common Order Ratio	$\frac{\text{Common Order Count}}{\text{Unique Order Count}}$

Table 5: Order Variability Features and Descriptions

Order Grouping Results

The authors grouped orders by terminology concept, encounter, service department, and DRG code. Grouping by terminology concept ensures that synonymous orders are counted together. The encounter level grouping is averaged over the length of stay. Grouping by encounter eliminates the dimensionality of time, which is generally difficult to model. Even with redundant clinical concepts grouped together, there were tens of thousands of unique UMLS concepts. The authors further reduced the dimensionality of orders by grouping by the service department of the order. This grouping resulted in lab tests such as blood amylase, basic metabolic panels, and tests for genetic markers all to be grouped together under the descriptor of “LAB.” Similarly procedures such as continuously pressurized ventilator support and pulmonary function exams were grouped under the descriptor of “Respiratory.” The authors performed this grouping to cope with a limited sample size of observations, and the large number of different orders. Mathematical models tend to break down if the number of predictors approaches or exceeds the number of observations. Lastly, patients with different diseases have different needs; grouping by DRG increases the comparability of patients within the same DRG. Grouping by DRG also scales the unit of analysis to a level familiar to hospital administrators. Figure 8 visualizes how the unit of analysis changed with each grouping criterion.

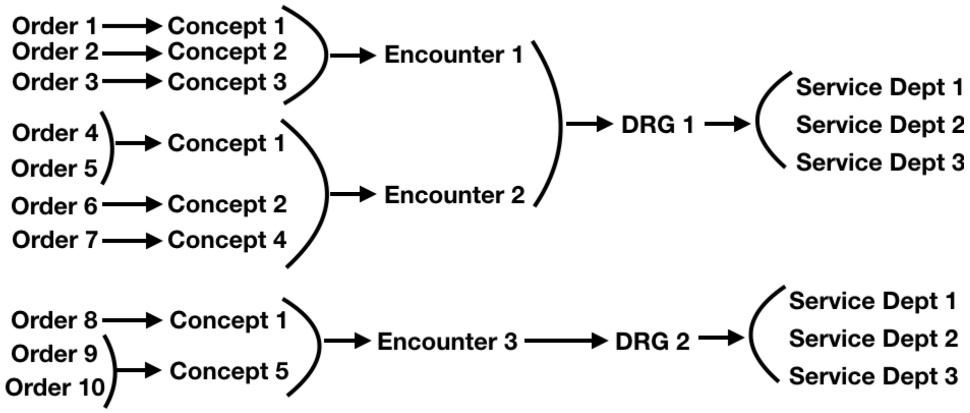


Figure 8: From Orders to Service Department Variance By DRG

Discussion & Limitations

The authors demonstrated the feasibility of processing orders from their raw form to a refined feature set at scale. The Vantage system accomplishes two high value tasks. The first being that it maps orders to a higher-level terminology. For the majority of orders the applied terminologies are part of the UMLS. The authors were able to achieve fairly accurate mappings without machine learning based methods. The results speak to the tractability of the problem, especially given the limited resources and generalizable methods used by the authors.

The second high value task accomplished by the Vantage System is that it estimates the number of nursing order completions over time. Author MCL is not aware of other research systems that have attempted this. Most other research systems treat nursing orders as singular entities or discard them[76–78,98].

Based on the total admissions between 7/1/2013 and 12/31/2016 Figure 3, VUMC experiences 206,908 encounters/30 months= 6,897 encounters/month. That figure is a little over half of the size of the project cohort. The demonstrated runtime of the system means that the system could incrementally process all orders monthly at a scale as large as VUMC and complete in less than a day. The Vantage system is executable with a single terminal command. The Vantage system was designed to dynamically update date parameters, enabling unsupervised incremental processing over time. If given an active connection to the data repository for orders and encounters, the Vantage system could run incrementally based on scheduled Cron job[99].

The authors' approach has several limitations. The Vantage system was built specifically for the WizOrder data model and it is not readily generalizable to other data models. That said, the EDW representation of the WizOrder data model is highly unstructured. While the system does not directly generalize, the method of construction might. The greatest limitation is that the authors' direct mapping strategy is not readily applicable to additional WizOrder data outside of the cohort. The specificity of the description mapping to the WizOrder system is an area for future work. The authors believe that nearest-neighbor clustering could provide a generalizable solution. The orders from the current cohort should be ample training data to fit such a model. Furthermore, the author only explored features they believed would be explanatory for order variability. The variables they created were not exhaustive of the possibilities and they focused on one particular task. The last major limitation is that the authors did not differentiate the intended number of times an order should be completed from the actual number of times an order was completed. In future work, the authors hope to incorporate data sources such as the medical administration record (MAR) along with lab and imaging systems to bridge this gap.

The Vantage system produced six different features for each of the sixteen different service departments in an attempt to quantify the variability care. The resulting 96 variables incorporate intended completions as well as conceptual groupings. The authors later selected and aggregated the features, and then evaluated the correlation of those features to the LOS ratio. The following chapter details how the authors used the derived order-based features to assess care variability.

Chapter IV

MODELING CARE VARIABILITY

Overview

In this segment of the study, the authors use the LOS ratio—the ratio of a patient's actual length of stay divided by the CMS calculated geometric average length of stay (of the same DRG) as the outcome of interest. The CMS average length of stay for a DRG is known as the expected length of stay[100].

Materials

The authors' care variability analysis used a 2015 MacBook Pro with four 2.9 GHz Intel processors and eight GB of RAM and a Linux server with twenty-four Xeon 2.00 GHz cores and 132 GB of RAM. The VUMC Finance office provided encounter-level cost data. Author MCL performed all statistical analysis with STATA version 29 Jan 2018.

Methods

Using a list of medical record numbers and encounter identifiers from the Chapter II cohort dataset, the authors requested cohort-related cost data from the VUMC Finance Department. The VUMC Finance office's methods for deriving cost are proprietary and are unknown to the authors. The Finance Department uses cost data operationally; this suggests that they were carefully derived.

The authors used a regression framework for their evaluation of order variability as a measure for care variability. The variance of the LOS ratio was the outcome of interest and each observation signified a DRG. DRGs were eliminated from consideration if they did not have at least ten encounters to average. The authors varied the minimum encounter threshold to twenty encounters and thirty encounters to evaluate the sensitivity of the results to this parameter. The authors compared a robust regression model with an order variability metric (built from the features extracted in Chapter III) to a robust regression model with cost variability. Each model used the same covariates, where only the variance term differs. The authors compared the fit of the models based on the Akaike information criterion (AIC) of both models. This criterion describes the information loss of a model based on the data. The model with the smallest AIC is relatively better than the other models. A difference in AIC of six or more suggests there is less than a 5% chance that the greater AIC model minimizes information loss better than the lesser AIC model. The study also reported the Student's T-Test of significance on the variable coefficients. The analysis was done at the DRG scale to develop a unified framework for evaluating care variability that could be used at the health system level by healthcare administrators or quality officers. In this framework, one DRG was compared to another. Comparing the LOS ratio between DRGs requires some form of case-mix adjustment, because the length of stay is sensitive to patient specific factors.

The authors chose covariates for their care variability analysis based upon a review of the literature of factors that affect the length of stay, and the availability of those factors in the data. Each model was adjusted for the intra-DRG variance of: age, sex, race, admission season, weekday admission indicator, Charlson Comorbidity Index[101], admission service, intensive care utilization indicator, and EPSI index. Any missing data was imputed using the aregImpute function from the Hmisc package.

To select which order features comprised the order variability measure, the authors used a scatter plot analysis. The scatter plot analysis involved plotting a single order feature on the X-axis and the outcome on the Y-axis. The authors combined the order features into one order variability measure using an average to permit a fairer comparison to cost variability. In statistical models, multiple predictors have a natural advantage over one predictor in terms of model fit. The regression methodology minimizes prediction error, which means that adding predictors cannot decrease the fit of a model.

Additional predictors do not decrease model fit because uncorrelated predictors can always be weighted to have zero effect. The three different order metrics were on different numeric scales. To weight each order-based feature equally, the authors centered and standardized each order-based feature before averaging them into a singular measure of order variability.

The authors performed two sub-analyses. The first eliminated encounters involving intensive/critical care to determine if the Vantage system could predict variability in more homogenous encounter groupings. The other sub-analysis examined if order variability could significantly predict unplanned readmissions (as defined by the 4th version of the CMS definition of unplanned readmissions). The variance of the length of stay was used as a covariate in the readmissions modeling sub-analysis.

Lastly, the authors purposefully and non-randomly chose a DRG with both a large number of encounters and a high amount of order variability. This analysis excluded patients that received intensive care during their admission. Next, the authors sampled 25 encounters from 25 unique patients coded with the chosen DRG for review. Author MCL examined the orders issued during each patient encounter and attempted to identify potential variations in care factoring in the patient's clinical condition. These findings were discussed with internist author CGW. This effort attempted to assess how the order variability metric might be used to prioritize DRGs for interventions such as standardizing approaches to care. The authors loosely explored if the order variability metric could directly inform potential sources of care variability within the sample.

Results

Figure 9 illustrates the total cohort size and the sub-analysis sample size. The sub-analysis excluding ICU patients further reduced the cohort by nearly a quarter.

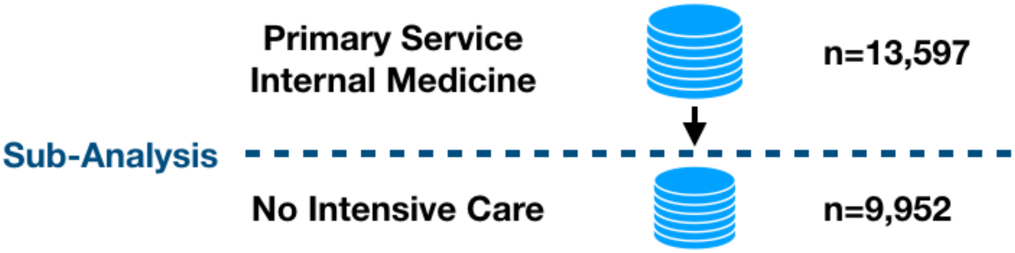


Figure 9: Cohort Sample Size with and without Sub-Analysis

The characteristics of the full study cohort and all covariates can be seen in Table 6. The cohort consists of encounters and not patients, meaning there may be multiple encounters for the same patient.

Variable	
Unique encounters (count)	13,597
Unique patients (count)	10,081
Age at admission	44.9 ± 13.1 years
Charlson Comorbidity Index	4.1 ± 1.5
Female sex	50.6%
Received intensive care	23.2%
Weekday Admission	75.5%
Race	
White	73.3%
Black	22.6%
Asian	1.1%
Unknown or other	3.0%
Season of Admission	
Spring (Mar-May)	21.9%
Summer (June-Aug)	26.8%
Fall (Sep-Nov)	27.9%
Winter (Dec-Feb)	23.4%
Admission Service	
Internal Medicine	74.6%
Emergency Room	3.3%
Infectious Disease	1.0%
General Surgery	1.0%
Others Not Listed	20.1%

Table 6: Care Variability Cohort Characteristics

The scatter plot analysis suggested that the number of commonly executed lab procedures and the total number of unique rehab orders were the most correlated order features. The authors scaled and centered the order features so that they could be averaged to form order variability. Table 7 shows that ranking DRGs by order variability produced different priorities compared to ranking by cost variability.

Variability Ranked DRG's	Number of Encounters	Variance of LOS Ratio
Order Variability		
1. Trach w/ ventilator support > 96hr	18	1.53
2. Septicemia or severe sepsis w/ ventilator support > 96hr	41	1.47
3. Septicemia or severe sepsis w/ complications	802	1.51
4. Pulmonary embolism w/ complications	44	1.74
5. Respiratory infections & inflammations w/ complications	97	1.50
6. Infectious/parasitic disease w/ operation & complications	136	1.44
7. Respiratory system diagnosis w/ ventilator support > 96hr	20	1.49
8. Respiratory system diagnosis w/ ventilator support < 96hr	28	1.01
9. Other kidney & urinary tract diagnoses w/ complications	112	1.75
10. Wound debridement and skin graft w/ complications	182	2.03
Cost Variability		
1. Osteomyelitis with complications	30	1.45
2. Alcohol/drug abuse or dependence w/o rehabilitation	295	0.90
3. Fever	40	1.12
4. Other respiratory system operation w/ complications	37	1.85
5. Intracranial hemorrhage or cerebral infarction w/ Complications	23	1.28
6. Gastrointestinal obstruction w/ complications	35	1.16
7. Other skin/subcutaneous tissue/breast procedures w/ Complications	26	1.59
8. Inflammatory bowel disease w/ complications	76	0.95
9. Inflammatory bowel disease	32	1.09
10. Cardiac arrhythmia and conduction disorders w/ complications	19	1.43

Table 7: Top 10 DRG's When Sorted by Variability

Table 8 presents the AIC, adjusted R^2 values, coefficient sign, and P–value for both order variability metric (order var) as well as cost variability (cost var) across different encounter thresholds. Order variability consistently produced better-fit models than cost variability across all thresholds for the minimum number of encounters. This result held in the sub-group without ICU patients.

	DRGs with 10+ Encounters	DRGs with 20+ Encounters	DRGs with 30+ Encounters
Full Cohort			
# of Encounters	12,598	11,391	10,574
# of DRGs	226	138	104
Adjusted-R^2			
Order Var	0.270	0.233	0.381
Cost Var	0.230	0.206	0.338
AIC			
Order Var	181.02*	73.23	0.40*
Cost Var	192.84	77.92	7.48
Coefficient Sign			
Order Var	+	+	+
Cost Var	+	+	-
Coefficient P-Value			
Order Var	0.001*	0.070	0.043*
Cost Var	0.224	0.245	0.605
Non-ICU Cohort			
# of Encounters	8,893	7,913	7,048
# of DRGs	188	114	79
Adjusted-R^2			
Order Var	0.180	0.225	0.244
Cost Var	0.139	0.204	0.176
AIC			
Order Var	136.20*	46.76	4.16*
Cost Var	145.54	49.87	12.63
Coefficient Sign			
Order Var	+	+	+
Cost Var	+	+	-
Coefficient P-Value			
Order Var	0.002*	0.018*	0.111
Cost Var	0.346	0.010*	0.823

Table 8: Care Variability Analysis Results with LOS Ratio as the Outcome

*** signifies statistical significance**

+ signifies that the LOS ratio variance is positively associated with that variable

- signifies that the LOS ratio variance is negatively associated with that variable

After running the regression, the authors used diagnostic methods to validate the assumptions of their analysis. A sample diagnostic (Residuals Versus Fitted values (RVF)) plot used by the authors can be seen in Figure 10. The RVF plot can visually indicate the presence of bias in the model fit and/or non-constant variance. The authors expected a roughly ellipse-shaped zero-centered pattern in the residual values, as demonstrated by the pattern below.

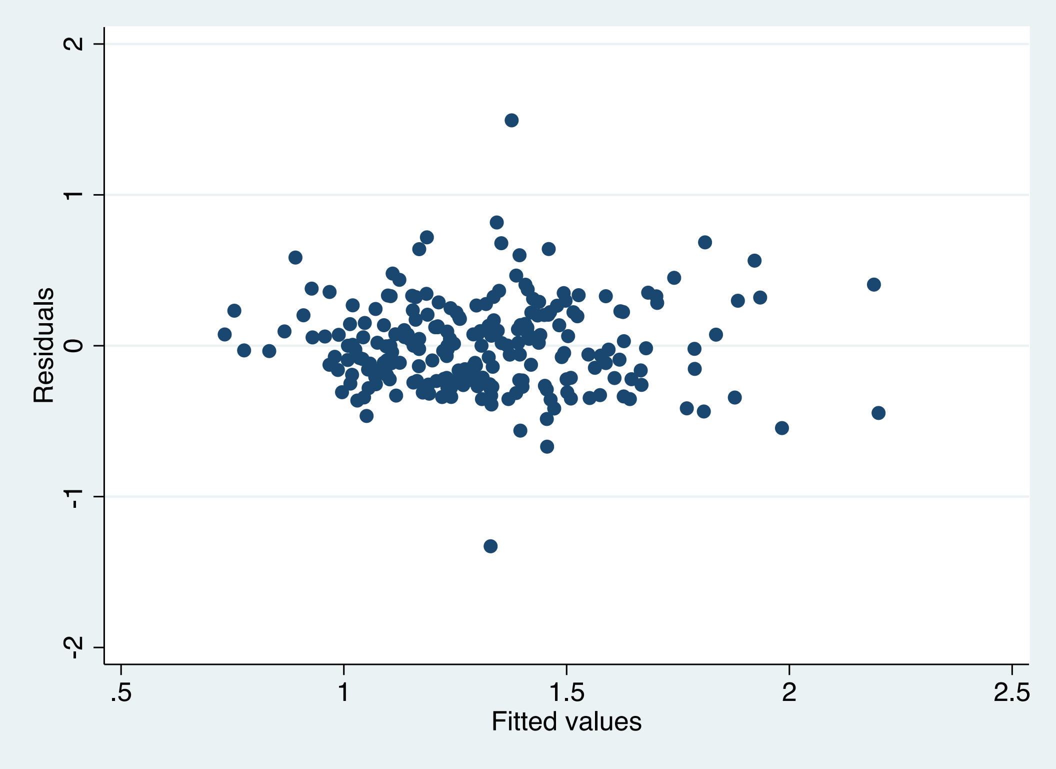


Figure 10: RVF Plot from LOS Ratio Analysis

The authors' order variability metric did not add significant value when modeling the variance of the unplanned readmission rate for a DRG compared to cost variability. This analysis included an additional covariate of the variance of the length of stay. The results are shown in Table 9.

	DRGs with 10+ Encounters	DRGs with 20+ Encounters	DRGs with 30+ Encounters
Full Cohort			
# of Encounters	12,598	11,391	10,574
# of DRGs	226	138	104
Adjusted-R^2			
Order Var	0.203	0.270	0.282
Cost Var	0.200	0.252	0.258

Table 9: Care Variability Analysis Results with Readmission Rate as Outcome

The authors explored if their care variability metric could inform quality improvement efforts in one DRG. The authors selected DRG 871 (Septicemia or severe sepsis w/ Complications) for review. DRG 871 had both a high order variability rank and prevalence in VUMC's case mix. This DRG was primarily composed of patients admitted with infectious disease diagnoses. The authors looked for patterns of care by chronologically looking at the orders of each patient in the sample. Generally the first 1-3 days were the most intensive for the patients in the sample. This meant that, the longer a patient stayed in the hospital, the fewer orders/day and nursing interactions per day they received. Table 10 demonstrates a comparative analysis using order features of Sepsis patients with a longer than expected length of stay and Sepsis patients with a shorter than expected length of stay. The table demonstrates that patients with longer lengths of stay have lower average order intensities, because the days toward the end of the encounter skew the average downward. One can see signs of greater variability of care in the high LOS patients. The total number of unique orders is always greater in the high LOS group, despite the greater number of encounter in the low LOS group. Author MCL also observed that the high LOS patients tended to have a higher number of commonly ordered laboratory tests. Patients that required in house rehabilitation services also generally had longer lengths of stays than those that did not use those services. The outcomes of the high LOS group were more variable than the low LOS group.

Feature	High LoS	Low LoS
Total Encounters	143	239
Median # of Labs/Day	3.1	8.9
Variance of Labs/Day	5.7	38.0
Total Unique Labs	328	311
Median # of Nursing Actions/Day	19.4	28.3
Variance of Nursing Actions /Day	90.2	152.5
Total Unique Nursing Orders	737	595
Median # of Medications/Day	7.6	21.6
Variance of Meds/Day	247.6	303.3
Total Unique Meds	1258	971

Table 10: Sample Analysis Using Order Features for Sepsis Patients

Discussion and Limitations

The authors tested their order variability metric across all eligible DRGs in VUMC's case mix. They found that order variability outperformed cost variability's ability to explain the LOS ratio. This result held throughout the sensitivity analysis of sample size, and in the sub-analysis excluding ICU patients. There were differences between the disorders prioritized by cost variability and order variability, as the two metrics only overlapped two DRG's out of the first ten listed. The coefficient of order variability remained consistently positive throughout all these analyses. The coefficient for cost variability did not consistently keep the same sign. It switched signs from positive to negative when the minimum number of encounters was increased to thirty. This inconsistency suggests that cost variability may have an inconsistent correlation with variability in patient outcomes. The interpretation of the order variability coefficient was that greater order variability contributed to greater deviations from the expected LOS ratio after accounting for the selected confounders. Order variability is useful as a surrogate to the LOS ratio, because order variability can be derived during the admission, while the LOS ratio cannot.

The order level interpretation is that a group of patients who more regularly ordered laboratory tests in common and who use a variety of rehabilitation services tend to have greater variability in outcomes, after adjusting for factors such as the number and severity of comorbidities and severity of illness. This finding also played out in the case of Sepsis patients. This finding in Sepsis patients may speak more to sepsis as a complex disease process, than the utility of the order variability metric.

The analysis was done at the level of an entire health system and across a wide variety of DRGs. This scale is a strength of the work. The authors have proposed and validated an order variability metric that is applicable to all DRG's. The authors used novel feature sets to attempt to identify variability in care. When modeling care variability, the authors corrected for many covariates (such as severity of illness, admission service, and an indicator for weekday admissions) that are often left out of analyses.

In future work the authors hope to incorporate admission shift (day versus night) and additional social factors. The admission shift data may become available for research with VUMC's new EMR e-Star. The authors plan to tap into the Behavioral Risk Factor Surveillance Survey using census tracts from the Center for Disease Control and Prevention's 500 Cities initiative[102]. A few select laboratory results from admission could have provided a more robust severity of illness measure. The authors hope that their approach can eventually inform services or care pathways to prioritize for standardization or improvement with an order-based measure of variability.

The primary limitation of this work is that the order variability metric is not prescriptive. The metric does not easily inform solutions for improving the case-mix adjusted LOS ratio. The order variability metric might identify potential opportunities for improvement. Further exploration is needed to determine the utility of this metric. In any case, chart review and observational studies would still play a role in devising implementable solutions. Next, this study used data from a single site, and further studies are needed to evaluate the generalizability of the authors' findings. The authors grouped patients by DRG because CMS calculates the expected length of stay for DRGs. There are other means of grouping patients and it is a future direction of this work to examine if the results change based on how patients are grouped. The authors used the EPSI index as a crude surrogate for severity of illness. They could have calculated one of the APACHE scoring systems, but were concerned about the degrees of freedom of their statistical model[103]. The fit of the models suggest that there are missing covariates. Because the model is under-fit, it is possible that the statistical importance of order variability is overstated. However, the model diagnostics do not suggest bias or violations of key regression assumptions.

Chapter V

CONCLUSIONS

This project examined the variability of care among patients with different DRGs through a novel point of view—orders. Previous research focused on cost, case-mix, and utilization differences between DRG's[12,45,104]. Orders can elucidate care that current cost capture mechanisms do not handle well, such as nursing utilization. To this end, the authors demonstrated the feasibility of a high throughput pipeline for order categorization based on terminologies from the UMLS. The study achieved reasonable accuracy with their concept mapping using open source software and terminologies, a single conceptual coder, and two reviewers. The authors could further augment their pipeline by including additional nursing concepts. The authors were able to map approximately 60% of all nursing orders to a UMLS concept. The extensive mapping efforts of the authors are not presently generalizable.

The authors further developed their methodological library with an imputation package. This package proved inferior to the Hmisc package for relatively simple datasets. The authors found that neither package had a statistically significant effect on the results of the care variability analysis.

The authors have explored the notion of care variability and its challenges. The authors developed unique order features that are potentially generally useful for length of stay predictive models. The features were systematically pared down to form a generalizable order variability measure. In a small pilot analysis, the authors' Vantage system bested the current standard metric (cost variability) for quality improvement when modeling patient outcomes. While additional, larger studies must independently validate this result, the authors preformed their analysis with a robust statistical method and assessed model dependencies. The authors' metric does not truly address the core issue of care variability. Much additional work must occur to capture the full context of decisions. Without that context, comparing care choices to guidelines, policies, and best practices is difficult.

REFERENCES

1. Kime, C. Patterns of Inpatient Care for Newly Diagnosed Immune Thrombocytopenia in US Children' s Hospitals. *Pediatrics* **131,** 880 (2013).

2. Macias, C. G. *et al.* Variability in inpatient management of children hospitalized with bronchiolitis. *Acad. Pediatr.* **15,** 69–76 (2015).

3. Rand, C. S., Powe, N. R., Wu, A. W. & Wilson, M. H. Why Don't Physicians Follow Clinical Practice Guidelines? *JAMA J. Am. Med. Assoc.* **Vol 282,** 1458–1465 (1999).

4. Francke, A., Smit, M., de Veer, A. & Mistiaen, P. Factors influencing the implementation of clinical guidelines for health care professionals: a systematic meta-review. *BMC med Inf. decis mak* **8,** 38 (2008).

5. Rozich, J. D. *et al.* Standardization as a mechanism to improve safety in health care. *Jt. Comm. J. Qual. Saf.* **30,** 5–14 (2004).

6. Bates, D. D. W., Boyle, D. L., Vliet, M. B. Vander, Schneider, J. & Leape, L. Relationship between medication errors and adverse drug events. *J. Gen. Intern. Med.* **10,** 199–205 (1995).

7. Ross, M. A. *et al.* Protocol-driven emergency department observation units offer savings, shorter stays, and reduced admissions. *Health Aff.* **32,** 2149–2156 (2013).

8. Jacke, C. O., Albert, U. S. & Kalder, M. The adherence paradox: guideline deviations contribute to the increased 5-year survival of breast cancer patients. *BMC Cancer* **15,** 734 (2015).

9. Britton, D. J., Bloch, R. B., Strout, T. D. & Baumann, M. R. Impact of a computerized order set on adherence to centers for disease control guidelines for the treatment of victims of sexual assault. *J. Emerg. Med.* **44,** 528–535 (2013).

10. Cisse, B. *et al.* Impact of socio-economic status on unplanned readmission following injury: A multicenter cohort study. *Injury* **47,** 1083–1090 (2016).

11. Kansagara, D. *et al.* Risk prediction models for hospital readmission: a systematic review. *JAMA* **306,** 1688–1698 (2011).

12. Williams, L. Modified DRGs as Evidence for Variability in Patient Severity. *Med. Care* **26,** 53–61 (1988).

13. Donzé, J., Lipsitz, S., Bates, D. W. & Schnipper, J. L. Causes and patterns of readmissions in patients with common comorbidities: retrospective cohort study. *BMJ* **347,** f7171 (2013).

14. Institute of Medicine. Crossing the quality chasm: a new health system for the 21th century. *Inst. Med.* 1–8 (2001). doi:10.17226/10027

15. Maas, J., Verheij, R. A., Groenewegen, P. P., de Vries, S. & Spreeuwenberg, P. Green space, urbanity, and health: how strong is the relation? *J Epidemiol Community Heal.* **60,** 587–592 (2006).

16. Betihavas, V. *et al.* What are the factors in risk prediction models for rehospitalisation for adults with chronic heart failure? *Aust. Crit. Care* **25,** 31–40 (2012).

17. Shahar, Y. & Musen, M. Plan Recognition and Revision in Support of Guideline-Based Care. *Work. notes AAAI Spring Symp. ...* 118–126 (1995).

18. James, B. C. & Savitz, L. A. How Intermountain Trimmed Health Care Costs Through Robust Quality Improvement Efforts. *Health Aff.* **30,** 1185–1191 (2011).

19. Rouse, W. B. & Cortese, D. A. *Engineering the system of healthcare delivery*. **153,** (IOS Press, 2010).

20. Yarbrough, P. M., Kukhareva, P. V, Spivak, E. S., Hopkins, C. & Kawamoto, K. Evidence-based care pathway for cellulitis improves process, clinical, and cost outcomes. *J. Hosp. Med.* **10,** 780–786 (2015).

21. Miller, P. L., Frawley, S. J. & Sayward, F. G. Maintaining and Incrementally Revalidating a Computer-Based Clinical Guideline: A Case Study. *J. Biomed. Inform.* **34,** 99–111 (2001).

22. Lobach, D. F. & Hammond, W. E. Computerized Decision Support Based on a Clinical Practice Guideline Improves Compliance with Care Standards. *Am. J. Med.* **102,** 89–98 (1997).

23. Timmermans, S. & Mauck, A. The Promises And Pitfalls Of Evidence-Based Medicine. *Health Aff.* **24,** 18–28 (2005).

24. Greenhalgh, T., Howick, J. & Maskrey, N. Evidence based medicine: a movement in crisis? *BMJ* **348,** 3725 (2014).

25. Teich, J. M. *et al.* Effects of computerized physician order entry on prescribing practices. *Arch. Intern. Med.* **160,** 2741–2747 (2000).

26. Neilson, E. G. *et al.* The impact of peer management on test-ordering behavior. *Ann. Intern. Med.* **141,**

196–204 (2004).

27. Peleg, M. *et al.* Comparing Guideline Models : A Case-study Approach. *J. Am. Med. Informatics Assoc.* **10,** 52–68 (2003).

28. Sanders, J. O., Bozic, K. J., Glassman, S. D., Jevsevar, D. S. & Weber, K. L. Clinical Practice Guidelines: Their Use, Misuse, and Future Directions. *J. Am. Acad. Orthop. Surg.* **22,** 135–144 (2014).

29. Hommersom, A., Groot, P. & Balser, M. in *Computer-based Medical Guidelines and Protocols: A Primer and Current Trends* 63–80 (2008).

30. Cassel, C. K. & Guest, J. A. Choosing Wisely: Helping Physicians and Patients Make Smart Decisions About Their Care. *JAMA* **307,** 1801–1802 (2017).

31. Bhatia, R. S. *et al.* Measuring the effect of Choosing Wisely : an integrated framework to assess campaign impact on low-value care. *BMJ Qual. Saf.* **24,** 523–531 (2015).

32. Safran, C. & Phillips, R. S. Interventions to Prevent Readmission : The Constraints of Cost and Efficacy. *Med. Care* **27,** 204–211 (1989).

33. J., O. K. *et al.* Measuring Diagnoses: ICD Code Accuracy. *Health Serv. Res.* **40,** 1620–1639 (2005).

34. Stein, D. J., Lund, C. & Nesse, R. M. Classification Systems in Psychiatry: Diagnosis and Global Mental Health in the Era of DSM-5 and ICD-11. *Curr Opin Psychiatry* **19,** 389–399 (2009).

35. SA, F. & Saint, S. Measuring pneumonia-related mortality using administrative data: Coding and consequences. *Ann. Intern. Med.* **160,** 430–431 (2014).

36. MB, R., PS, P., Priya, A. & PK, L. Variation in diagnostic coding of patients with pneumonia and its association with hospital risk-standardized mortality rates: A cross-sectional analysis. *Ann. Intern. Med.* **160,** 380–388 (2014).

37. Burns, E. M. *et al.* Systematic review of discharge coding accuracy. *J. Public Health (Bangkok).* **34,** 138–148 (2012).

38. Tehrani, S. *et al.* Fibrin clot properties and haemostatic function in men and women with type 1 diabetes. *Thromb. Haemost.* **113,** 312–318 (2015).

39. Hess, K. The vulnerable blood. Coagulation and clot structure in diabetes mellitus. *Hamostaseologie* **35,**

25–33 (2015).

40. Udvardy, M., Posan, E. & Harsfalvi, J. Altered lysis resistance of platelet-rich clots in patients with insulin-dependent diabetes mellitus. *Thromb. Res.* **79,** 57–63 (1995).

41. Quinn, K. After the revolution: Drgs at age 30. *Ann. Intern. Med.* **160,** 426–429 (2014).

42. Eldenburg, L. & Kallapur, S. Changes in hospital service mix and cost allocations in response to changes in Medicare reimbursement schemes. *J. Account. Econ.* **23,** 31–51 (1997).

43. Taheri, P. a, Butz, D. a & Greenfield, L. J. Length of stay has minimal impact on the cost of hospital admission. *J. Am. Coll. Surg.* **191,** 123–130 (2000).

44. Horn, S. D., Sharkey, P. D., Chambers, A. F. & Horn, R. A. Severity of illness within DRGs: Impact on prospective payment. *Am. J. Public Health* **75,** 1195–1199 (1985).

45. Smits, H. L., Fetter, R. B. & McMahon, L. F. Variation in resource use within diagnosis-related groups: The severity issue. *Health Care Financing Review* **1984,** 71–78 (1984).

46. Williams, L. Variation in Resource Use within Diagnosis-Related Groups : The Effect of Severity of Illness and Physician Practice. *Med. Care* **24,** 388–397 (1986).

47. Thompson, J. D. The measurement of nursing intensity. *Health care financing review* **Suppl,** 47–55 (1984).

48. Pirson, M. *et al.* Variability of nursing care by APR-DRG and by severity of illness in a sample of nine Belgian hospitals. *BMC Nurs.* **12,** 26 (2013).

49. Pirson, M. *et al.* Variability of nursing care by APR-DRG and by severity of illness in a sample of nine Belgian hospitals. *BMC Nurs.* **12,** 26 (2013).

50. Knauf, R. A., Ballard, K., Mossman, P. N. & Lichtig, L. K. Nursing Cost by DRG: Nursing Intensity Weights. *Policy, Polit. Nurs. Pract.* **7,** 281–289 (2006).

51. Welton, J. M., Fischer, M. H., DeGrace, S. & Zone-Smith, L. Hospital nursing costs, billing, and reimbursement. *Nurs Econ* **24,** 227,239-245,262 (2006).

52. Abbass, I. Variability in the Initial Costs of Care and One-Year Outcomes of Observation Services. *West. J. Emerg. Med.* **16,** 395–400 (2015).

53. Shwartz, M., Young, D. W. & Siegrist, R. The Ratio of Costs to Charges: How Good a Basis for Estimating Costs? *Inquiry* **32,** 476–481 (1995).

54. Schwartz, A. L., Chernew, M. E., Landon, B. E. & McWilliams, J. M. Changes in Low-Value Services in Year 1 of the Medicare Pioneer Accountable Care Organization Program. *JAMA Intern. Med.* **02115,** 1–11 (2015).

55. Borghans, I., Heijink, R., Kool, T., Lagoe, R. J. & Westert, G. P. Benchmarking and reducing length of stay in Dutch hospitals. *BMC Health Serv. Res.* **8,** 220 (2008).

56. Peterson, E. D. *et al.* Hospital variability in length of stay after coronary artery bypass surgery: Results from the Society of Thoracic Surgeon's National Cardiac Database. *Ann. Thorac. Surg.* **74,** 464–473 (2002).

57. Horn, S. D. ., Horn, R. A. ., Sharkey, P. D. . & Chambers, A. F. . Severity of Illness within DRGs : Homogeneity Study. *Med. Care* **24,** 225–235 (1986).

58. McMahon, L. F. & Newbold, R. Variation in Resource Use within Diagnosis-Related Groups : The Effect of Severity of Illness and Physician Practice. *Med. Care* **24,** 388–397 (1986).

59. Dimick, J. B., Pronovost, P. J., Heitmiller, R. F. & Lipsett, P. A. Intensive care unit physician staffing is associated with decreased length of stay, hospital cost, and complications after esophageal resection. *Crit. CARE Med.* **29,** 753–758 (2001).

60. Riguzzi, C., Hern, H. G., Vahidnia, F., Herring, A. & Alter, H. The july effect: is emergency department length of stay greater at the beginning of the hospital academic year? *West. J. Emerg. Med.* **15,** 88–93 (2014).

61. Singh, S. & Fletcher, K. E. A qualitative evaluation of geographical localization of hospitalists: How unintended consequences may impact quality. *J. Gen. Intern. Med.* **29,** 1009–1016 (2014).

62. Efron, B. Missing Data, Imputation, and the Bootstrap. *J. Am. Stat. Assoc.* **89,** 463–475 (1994).

63. Sterne, J. A. C. *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* **338,** (2009).

64. LANDERMAN, L. R., LAND, K. C. & PIEPER, C. F. An Empirical Evaluation of the Predictive Mean

Matching Method for Imputing Missing Values. *Sociol. Methods Res.* **26,** 3–33 (1997).

65. Gerko, V., E., F. L., Jeroen, P. & Stef, B. Predictive mean matching imputation of semicontinuous variables. *Stat. Neerl.* **68,** 61–90 (2014).

66. Harrell, J. F. E. Package 'Hmisc'. (2018).

67. Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Doove, L. & Jolani, S. Multivariate Imputation by Chained Equations: MICE V1.0 User´s manual. 121 (2000). doi:10.18637/jss.v045.i03>.

68. Berger, R. G. & Kichak, J. P. Computerized physician order entry: helpful or harmful? *J. Am. Med. Informatics Assoc.* **11,** 100–103 (2004).

69. Lindberg, D. A. B. Commentary on G . Octo Barnett's Report to the Computer Research Study Section. *J. Am. Med. Informatics Assoc.* **13,** 136–137 (2006).

70. Wentzer, H. S., Böttger, U. & Boye, N. Unintended transformations of clinical relations with a computerized physician order entry system. *Int. J. Med. Inform.* **76,** 456–461 (2007).

71. Adler-Milstein, J. *et al.* Electronic Health Record Adoption In US Hospitals: Progress Continues, But Challenges Persist. *Health Aff.* **34,** 2174–2180 (2015).

72. Ash, J. S. *et al.* A cross-site qualitative study of physician order entry. *J. Am. Med. Inform. Assoc.* **10,** 188–200 (2003).

73. Chen, J. H. & Altman, R. B. Mining for clinical expertise in (undocumented) order sets to power an order suggestion system. *AMIA Jt. Summits Transl. Sci. Proc.* **2013,** 34–8 (2013).

74. Leu, M. G., Morelli, S. A., Chung, O.-Y. & Radford, S. Systematic Update of Computerized Physician Order Entry Order Sets to Improve Quality of Care: A Case Study. *Pediatrics* **131,** S60–S67 (2013).

75. Chen, J. H. & Altman, R. B. Automated Physician Order Recommendations and Outcome Predictions by Data-Mining Electronic Medical Records. *AMIA Summits Transl. Sci. Proc.* 206–210 (2014).

76. Chen, J. H., Goldstein, M. K., Asch, S. M., Mackey, L. & Altman, R. B. Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets. *J. Am. Med. Informatics Assoc.* **24,** ocw136 (2016).

77. Klann, J. G., Szolovits, P., Downs, S. M. & Schadow, G. Decision support from local data: Creating

adaptive order menus from past clinician behavior. *J. Biomed. Inform.* **48,** 84–93 (2014).

78. Wright, A. & Sittig, D. F. Automated development of order sets and corollary orders by data mining in an ambulatory computerized physician order entry system. *AMIA Annu. Symp. Proc.* 819–23 (2006).

79. Cimino, J. J., Clayton, P. D., Hripcsak, G. & Johnson, S. B. Knowledge-based Approaches to the Maintenance of a Large Controlled Medical Terminology. *J Am Med Inf. Assoc* **1,** 35–50 (1994).

80. Jurisica, I., Mylopoulos, J. & Yu, E. Ontologies for knowledge management: an information systems perspective. *Knowl. Inf. Syst.* **6,** 380–401 (2004).

81. Sittig, D. F. *et al.* The state of the art in clinical knowledge management: An inventory of tools and techniques. *Int. J. Med. Inform.* **79,** 44–57 (2010).

82. Ihaka, R. & Gentleman, R. R: a language for data analysis and graphics. *J. Comput. Graph. Stat.* **5,** 299–314 (1996).

83. McKinney, W. *Pandas: a Foundational Python Library for Data Analysis and Statistics*. (2010).

84. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2012).

85. Horn, S. D., Horn, R. A. & Sharkey, P. D. The Severity of Illness Index as a severity adjustment to diagnosis-related groups. *Health care financing review* **Suppl,** 33–45 (1984).

86. Nash, W. J., Sellers, T. L., Talbot, S. R., Cawthorn, A. J. & Ford, W. B. The population biology of abalone (haliotis species) in Tasmania. I. Blacklip Abalone (h. rubra) from the north coast and islands of Bass Strait. *Sea Fish. Div. Tech. Rep.* (1994).

87. Walsh, C. G., Ribeiro, J. D. & Franklin, J. C. Predicting Risk of Suicide Attempts Over Time Through Machine Learning. *Clin. Psychol. Sci.* **5,** 457–469 (2017).

88. Geissbuhler, A. & Miller, R. A. WizOrder, a user-friendly interface for order entry and clinical decision support tools. in *Proceedings of the Annual Symposium on Computer Application in Medical Care* 1002 (American Medical Informatics Association, 1995).

89. Geissbühler, A. & Miller, R. A. A new approach to the implementation of direct care-provider order entry. in *Proceedings of the AMIA Annual Fall Symposium* 689 (American Medical Informatics Association, 1996).

90. Geissbuhler, A. & Miller, R. A. Distributing knowledge maintenance for clinical decision-support systems: the ‘knowledge library’ model. in *AMIA Annual Symposium* 770–774 (1999). at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2232510&tool=pmcentrez&rendertype=abstract>

91. Starmer, J. M., Talbert, D. A. & Miller, R. A. Experience using a programmable rules engine to implement a complex medical protocol during order entry. in *Proceedings of the AMIA Symposium* 829 (American Medical Informatics Association, 2000).

92. Aronson, A. R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. in *Proceedings of the AMIA Symposium* 17 (American Medical Informatics Association, 2001).

93. Xu, H. *et al.* MedEx: a medication information extraction system for clinical narratives. *J. Am. Med. Informatics Assoc.* **17,** 19–24 (2010).

94. Liu, S., Ma, W., Moore, R., Ganesan, V. & Nelson, S. RxNorm: prescription for electronic drug information exchange. *IT Prof.* **7,** 17–23 (2005).

95. Snomed, C. T. Systematized nomenclature of medicine-clinical terms. *Int. Heal. Terminol. Stand. Dev. Organ.* (2011).

96. *The MEDCIN Clinical Terminology*. (Medicom Systems).

97. *Healthcare Common Procedure Coding System (HCPCS)*. (Centers for Medicare & Medicaid Services, 2003).

98. Chen, J. H. & Altman, R. B. Data-Mining Electronic Medical Records for Clinical Order Recommendations: Wisdom of the Crowd or Tyranny of the Mob? *Arslan, A. K., Colak, C., Sarihan, M. E. (2016). Differ. Med. data Min. approaches based Predict. ischemic stroke. Comput. Methods Programs Biomed. 130, 87–92. http//doi.org/10.1016/j.cmpb.2016.03.022 Care, M. (2015). Ann.* **2015,** 435–9 (2015).

99. Open Group, T. The Open Group Base Specification. *IEEE* (2018).

100. CMS. Department of Health and Human Services Regulations. *Fed. Regist.* **78,** (2013).

101. Kastner, C. *et al.* The Charlson comorbidity score: a superior comorbidity assessment tool for the prostate cancer multidisciplinary meeting. *Prostate Cancer Prostatic Dis.* **9,** 270–274 (2006).

102. Prevention, C. for D. C. and. 500 Cities: Local data for better health. (2017). at <https://www.cdc.gov/500Cities/>

103. Knaus, W. A., Draper, E. A., Wagner, D. P. & Zimmerman, J. E. APACHE II: a severity of disease classification system. *Crit. Care Med.* **13,** 818—829 (1985).

104. Schreyögg, J., Tiemann, O. & Busse, R. Cost accounting to determine prices: How well do prices reflect costs in the German DRG-system? *Health Care Manag. Sci.* **9,** 269–279 (2006).