Measurement and Prediction of Inpatient Case Manager Workload in a Tertiary Hospital Setting

by

Jason Edward Stuck

B.S. Electrical Engineering, University of Wyoming, Laramie, 2007
B.A. International Studies, University of Wyoming, Laramie, 2007

Submitted to the Institute for Data, Systems, and Society and the MIT Sloan School of Management in partial fulfillment of the requirements for the degrees of

Master of Science in Engineering Systems

and

Master of Business Administration

in conjunction with the Leaders for Global Operations Program at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2016

Signature redacted

Author .....
Engineering Systems Division and the MIT Sloan School of Management
August 5, 2016

Signature redacted

Certified by .....
Retsef Levi, Thesis Supervisor
J. Spencer Standish Professor of Management, Professor of Operations Management
MIT Sloan School of Management

Signature redacted

Certified by .....
David Simchi-Levi, Thesis Supervisor
Professor of Engineering Systems, Institute for Data, Systems, and Society
Department of Civil Engineering

Signature redacted

Approved by .....
John N. Tsitsiklis
Clarence J. Lebel Professor of Electrical Engineering, IDSS Graduate Officer

Signature redacted

Approved by .....
Maura Herson
Director, MBA Program, MIT Sloan School of Management

THIS PAGE INTENTIONALLY LEFT BLANK

Measurement and Prediction of Inpatient Case Manager Workload in a Tertiary Hospital Setting

by

Jason Edward Stuck

Submitted to the Institute for Data, Systems, and Society and the MIT Sloan School of Management on August 5, 2016, in partial fulfillment of the requirements for the degrees of

Master of Science in Engineering Systems

and

Master of Business Administration

Abstract

A patient's care needs often extend past discharge from an acute hospital setting. At Massachusetts General Hospital (MGH), inpatient case managers, acting in a discharge planning capacity, help develop and coordinate the execution of plans, specifically tailored for a patient, to ensure these care needs are met. Case managers, and case management leadership, must confront multiple sources of workload variability across different time and scale perspectives.

Case managers are assigned a relatively invariant number of cases by floor. Inter-floor workload variability exists because the "typical" case on one floor may require more or less work than the "typical" case on another floor. Inter-case variability is also present; for a given case manager, the concept of a "typical" case has limited value. Some cases require essentially no work from a discharge planning case manager, while other cases can consume many hours, either on a single day, or spread across multiple days. The case characteristics determining the amount of work required of a case manager are not solely, or even primarily, clinical. Instead, discharge disposition, insurance considerations, patient preferences, and a wide array of psycho-social factors, as well as complex interactions among case characteristics, drive the workload for any single case. Finally, the total amount of work required, across all assigned cases, can vary dramatically from day to day.

In any discussion of case manager workload, variability, in all of its dimensions, is a fundamental characteristic. From an operational improvement standpoint workload variability has to be fully considered, understood, and accommodated. The current static staffing scheme, based on the number of beds a case manager is responsible for, does not adequately address the observed variability in daily workload. Therefore, the ultimate objective of our work is to develop a candidate staffing scheme and staffing guidelines incorporating requisite dynamic element to address variability in a case manager's daily workload and/or reduce observed upside variability.

Since the requisite understanding of workload variability will always prove elusive without a meaningful way to measure workload, in the first, necessary step for our work we develop a method of measuring the amount of work performed by a case manager, for a given case or on a given day. Though the scale for our work metric requires more refined calibration, it allows one to say with a high degree of certainty that "this case required more work than that case" or "this day represented a higher workload for a case manager than that day". The source of the score for a case or day is the work documented in case manager notes. We develop an automated scoring

procedure to retrospectively score cases based on the text of case manager notes. At the heart of our text-analytical engine is an augmented bag-of-words approach that preserves the relevant context for a case manager note. Using a regression tree to operate on our text feature vector for a case note results in validation set scoring with an R^2 of 0.98 at the case and day level.

In validating our scoring methodology case managers were asked to rank a group of cases in order of increasing workload. This ordinal ranking was compared to the ranking derived from our work score and yielded a value for Kendall's coefficient of concordance, W, of 0.98, indicating exceptional agreement.

Results using our score provide further indirect support for the validity of our scoring methodology. For example, the top decile of patients by work score accounted for 40% of the total work scored. This is in line with case manager reports that a relatively small number of patients require a disproportionately large amount of case manager time.

Our validated work score is then used as a response variable for explanatory and predictive modeling of case manager workload. The predictor variables are derived from a phased framework we developed over the course of our work. That is, distinct phases can be identified on a discharge planning plane as a patient progresses to ultimate discharge. For the majority of cases it is possible to identify, unambiguously, which phase a case is in. Counts of the number of cases in each phase at 04:00 form our predictor variables in projecting the amount of case manager workload required for the upcoming day.

Each phase is associated with both a characteristic amount of work and, as importantly, whether a given case will require any case manager work on a given day. This allows us to introduce the concept of an active census or active caseload. It is this concept that allows us to capture a key, under-considered source of variability - whether a case will require any work of a case manager on a given day. Using a regression-based model, the work for a case manager can currently be predicted with an R^2 of 0.51 and a case can be predicted as active with an R^2 of 0.66. With classification based on a boosted tree, a day can be correctly predicted as high, medium, or low workload with an accuracy of 81%. Two class misclassification error rates (high-as-low or low-as-high) of 7% can currently be achieved. Finally, in a synthesis of all of our work, we present the outline for a dynamic case assignment scheme based on pooling and balancing the number of cases in each phase between case managers within a pool. This can help attenuate the magnitude of high workload days and reduce upside variability.

Thesis Supervisor: Retsef Levi
Title: J. Spencer Standish Professor of Management, Professor of Operations Management
MIT Sloan School of Management

Thesis Supervisor: David Simchi-Levi
Title: Professor of Engineering Systems, Institute for Data, Systems, and Society
Department of Civil Engineering

Acknowledgments

Though it is unlikely that my faith in the power of individual human agency will ever be shaken, it should go without saying that this thesis is only possible as a product of the efforts of a great many people. Saying this in no way absolves me of responsibility for any deficits in this work. More to the point, it is safe to assume any noteworthy aspects of this thesis are the results of drawing upon the inspiration and advice of others.

Of course, the technical advice and overall direction provided by my esteemed advisors, Professors Retsef Levi and David Simchi-Levi was indispensable. Yet, it was other types of advice, related to maintaining perspective and balance that proved the most beneficial and, in many ways, prescient. It would be inappropriately self-indulgent to describe my personal trials over the preceding months but, suffice it to say, the most challenging period of my life was coincident with the work described in this thesis; this is true despite the fact that I have spent extended periods of time in inhospitable environments over the course of my Army career. The accessibility and proffered "words of wisdom" from my advisors helped me to surmount these challenges.

Cecilia Zenteno and Aleida Braaksma, in tirelessly providing suggestions for revisions, deserve the lion's share of credit for helping me communicate the important parts of our work in a way accessible to a larger audience of readers. Cecilia and Aleida also helped focus my efforts, no small task given my admitted penchant for wide-ranging intellectual curiosity at times when concentrating on the task at hand would serve better. Finally, to both of you, I owe thanks for helping me balance a tendency toward being unproductively self-critical.

I also wish to acknowledge and thank Dr. Peter Dunn and Bethany Daily. First, as part of the selection team for the MGH LGO internship, along with Retsef and Cecilia, I am grateful for the opportunity I had to work at MGH. Second, thank you for helping to provide the structure necessary to allow LGOs to leverage their skills and complete meaningful work in the complex environment at MGH. Without the collaboration that this structure facilitates our work would not have been possible. Similarly, your advice and guidance was key in developing a feasible path for our work.

This work would have been impossible without the support of case management leadership at MGH – Nancy Sullivan, Rachael McKenzie, and Debra Connolly. The commitment shown in always striving to improve case management at MGH was evident from the outset of our work; this commitment and your support was unwavering during the project. I hope the results of the work detailed in this thesis begin to repay the time and effort you invested in project success.

In a profoundly fundamental way it was the direct observation of case managers at MGH that inspired this work and, more specifically, the sheer amount of effort invested in this work. Given my own background, the tireless service I witnessed, as case managers worked on behalf of patients to ensure care needs were met, resonated deeply. Case managers, despite the often frenetic pace of their work, freely shared their time, experience, and insights to provide the context in which this work is grounded. Case managers I observed or interviewed directly included: Nora Arbeene, Wanda Quatrale, Sheila Cambra, Laurene Dynan, Alecia Laing, Maria Sweeney, Maria Seavey, Michael Trotta, Lorraine Zoda, Karen Hawko, Angela Wynder, Dana Madden, Margaret 'Peggy' Johnson, and Andrea Belliveau. Sometimes the interviews were scheduled, sometimes impromptu, but I thank all of you for your patience in answering my questions. I hope my presence was minimally disruptive and, again, I hope this work can form part of the canon of future work leading to improvements

that positively impact you. Dawn Cushing also deserves thanks for helping me gain access to the data necessary for the analysis at the heart of our work.

I would be remiss if I did not acknowledge and thank the LGO program. Until recent events I did not fully appreciate how special the community aspect of LGO is. In particular, I would like to thank Thomas Roemer, Ted Equi, and Patricia 'Patty' Eames. Without the continuing support of the LGO program this thesis would not exist.

Finally, I want to thank my wife, Stephanie. As stated, the support of all those named above was critical for this thesis. However, in the final analysis, even this support would have been insufficient without the support of my wife.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter serves as a brief introduction to the environment in which our work was performed, as well as providing an outline of the primary phases completed during the course of our work. The chapter ends with a brief, essentially qualitative, discussion of possible methodological extensions of our work, and associated analysis techniques, beyond the specific context of inpatient discharge planning at Massachusetts General Hospital (MGH).

From a high-level perspective the motivation for our work was simple. MGH's case managers, and case management leadership, report that the daily workload for case managers is subject to marked variability. Cursory objective indications, such as variability in the daily number of case manager notes, support this claim [1].

Our aims include a rigorous analysis of workload variability in order to develop operational improvements focused on staffing policies that incorporate a feasible dynamic component. In fact, workload variability, sometimes extreme, is an inherent feature of the essentially static nature of current staffing and case assignment policies. To explain, case managers are responsible for patients in an assigned pool of beds. Due to high capacity utilization at MGH this results in a relatively invariant daily case manager caseload; the number of occupied beds fluctuates in a very narrow range on a day-to-day basis. However, the number of cases that require work on a given day, the type of work for each case, and the amount of work for a case varies widely; caseload is not a useful indicator of workload on a daily basis.

The fact that caseload is not equivalent to workload is widely recognized, as is the fact that workload variability and unequal distribution of workload can have serious negative impacts on operations[13][12][95][38][105]. Though our work does not investigate the effects of workload variability on patient length-of-stay(LOS), it is possible that sub-optimality in daily case manager resource allocation (staffing) could introduce delays to patient discharge. What is certain is that workload variability can lead to job dissatisfaction and eventual burnout for frontline case managers. This state of affairs, and any resulting staff turnover, has negative impacts for the system.

In order to effectively grapple with workload variability, it is first necessary to develop a method to quantify workload. As other researchers have noted, there is no generally accepted system for

[1] A complete discussion of the utility and deficits of case manager note counts used as an indicator of workload is provided in Chapter 4

quantifying case manager workload in a tertiary inpatient setting like MGH[14][38][95]. Therefore we develop and present a methodology allowing automatic, retrospective scoring of the work completed for a case by a case manager. Our work score or metric is then used as a response variable in developing predictive models of the work that will be required for a set of cases in the succeeding day. Finally, we use the insights gained from our work to develop a candidate dynamic case assignment scheme and guidance for staffing policies.

To close out this introductory section and provide a better appreciation for the context in which our work was completed, it is worth considering other systems for quantifying workload in a healthcare setting. It may seem that a natural analog to our efforts are the various workload quantification systems used to inform staffing decisions for attending nurses; many of these types of systems exist with an extended use history and associated research efforts[98][89][33][46][47][64][103][66][83][100][105]. Ostensibly these are "objective" systems, but workload measures and staffing recommendations arising from different systems do not necessarily exhibit a high degree of correspondence.[48][84][104][100][42]

The reason for any disagreements among systems likely stems from an oft-used, but not necessarily well-defined attribute of a patient, provider, or system - acuity[45]. As Brennan and Daly point out, acuity has many attributes, such as physical, psychological, care needs, complexity, and workload. Our efforts focus on the use of work or workload as the base concept. The other attributes of acuity may or may not map to the work a case manager has to perform for a case.

1.1 Background

This section provides a brief overview of MGH, past MIT-MGH collaborative efforts, and inpatient case management at MGH[2].

1.1.1 Massachusetts General Hospital

By any number of metrics, Massachusetts General Hospital (MGH) is consistently ranked as one of the top hospitals in the United States [27]. In addition to serving as a teaching hospital, affiliated with Harvard Medical School, and providing primary care services for the community, MGH enjoys worldwide renown as a leading tertiary healthcare institution. In its mission to deliver care of the highest quality, MGH serves nearly 50,000 inpatients and 1.5 million outpatients annually [27]. In addition to being the largest hospital in New England, MGH is also a leading research center, conducting the largest hospital-based research program in the world [27].

1.1.2 MIT-MGH Collaboration

Spanning more than a decade, the history of the MIT-MGH Collaboration can aptly be described as storied, productive, and mutually beneficial. Beginning as a relationship between MGH and the Massachusetts Institute of Technology (MIT) Sloan School of Management, since 2011 MGH has also served as a partner organization with MIT's Leaders for Global Operations (LGO) program.

[2]Chapter 2 provides a more detailed discussion of case management, particularly in the context of MGH

The collaborative efforts are built upon a simple, yet powerful premise: striving to improve the operational efficiency and effectiveness at MGH while maintaining the highest standards of care for the patients served by MGH. The team involved in operational improvement initiatives comprises MGH leadership, MIT Faculty, MIT Sloan School of Management - Operations Management Group post-doctoral fellows, and LGO students completing a six-month research internship at MGH[29].

The scope of the collaboration has expanded since the earliest efforts to develop and implement improvements for perioperative processes, efforts that included bed allocation/assignment management[65] and surgical inpatient flow optimization[79][110][56][108][119]. Other recent efforts have been aimed at prescription management in a primary care setting[116] and rationalizing infusion clinic appointment scheduling procedures[111]. This is just a small sample of the collaborative efforts and all have helped improve patient access to the care MGH can provide. The structure provided for the MIT-MGH Collaboration helps augment the speed with which potential areas for improvement can be identified and investigated, thereby allowing any protracted process of implementation to begin that much sooner. The work described in this thesis marks the first foray into the realm of inpatient case management for the MIT-MGH Collaboration. It is hoped that this work can form a foundational part of a canon of future work in this area, not limited to MGH, primarily by revealing promising directions for further work.

1.1.3 Inpatient case management at MGH

Both because this is the first work in the area of case management for the MIT-MGH Collaboration, and because the very term case management may denote fundamentally different functions and processes depending on the setting, Chapter 2 provides an extensive description of inpatient case management at MGH. As an introduction, case managers primarily perform two critical roles at MGH: utilization review (UR) and discharge planning (DCP). In the UR role case managers help ensure that a patient, upon admission, requires the acute level of care that MGH can provide. As the patient progresses, from a clinical standpoint, UR case managers help to ensure that patients continue to warrant hospitalization. The UR role is important in helping to determine if the hospital's limited resources are being used to serve patients whose needs cannot be met more appropriately in a sub-acute setting. Further, UR case managers act as *de facto* patient advocates in coordination with payors by helping to make sure services rendered will be covered by payors. The target benchmark caseload for a UR case manager is 50 cases.

DCP case managers work in coordination with other members of the care team to ensure a patient's care needs are met after discharge from MGH. Often a patient, though not requiring the level of care that MGH is suited to provide, may have extensive needs extending past discharge day. DCP case managers balance myriad factors, including needs, preferences, insurance considerations, and psycho-social factors, to develop a safe discharge plan tailored specifically for a patient. The primary focus of this work is on the DCP function performed by inpatient case managers at MGH.

As with UR case managers, a typical benchmark caseload exists for DCP case managers - 25 cases. However, case managers on some floors may have fewer or more cases. This varying benchmark is in recognition of the fact that the "typical" case on one floor may require more DCP work than the "typical" case on another floor. For example, floors with a high percentage of patients undergoing elective procedures generally have well-characterized discharge needs and DCP paths. Similarly, insurance considerations, because of pre-authorizations associated with elective procedures, usually

do not factor as prominently into the workload for DCP case managers on these floors. In contrast, on floors associated with unplanned admissions, the DCP case manager must contend with emergent issues resulting from an interplay of patient/case characteristics and a patient's needs upon discharge. Depending on how these factors interact a case could present virtually no work for a case manager or a case could require many hours of a case manager's time, either on a single day or spread across multiple days.

The above touches on two sources of potential variability in a case manager's workload, despite a relatively invariant number of cases. The first of these can aptly be termed inter-floor variability and traces to a differing prevalence of certain case characteristics between floors. The second type of variability is inter-case variability where, for a given case manager, one case may consume much more time as compared to another case.

Another source of variability is inter-day variability. This variability can be traced to two primary sources. First, on a given day a case manager has multiple cases, each of which can be in a different stage of the discharge planning process; this mix changes daily. Second, not all assigned cases are active on a given day. Here active is used as a convention to encapsulate the fact that a case may be marked by latent periods when relatively little work is done for a case. This is natural as it takes time for a requisite degree of certainty to emerge about a patient's ultimate needs upon discharge. Until this modicum of clarity is achieved it is difficult to perform discharge planning for a case. Further, developing a discharge plan may require coordination with extra-MGH entities. This coordination is marked by delays as the outside entity performs its own review of a case.

Given the above discussion, it is difficult to determine definitively what an appropriate benchmark caseload is. Furthermore, variability, across many dimensions, is a fundamental characteristic in case manager workload. In order to make any operational improvements to DCP this variability must first be understood. This variability also suggests that a dynamic component for case assignment is necessary to accommodate any irreducible variability, the amount of which is still significant. Of course, all of this presupposes a way to accurately measure workload. This thesis details our efforts to first develop and validate a work score or metric. This metric is then used as a response variable for explanatory and predictive models, models that can be used to determine when a case manager may face an excessively high workload in the coming day. Finally, the work metric is used to help identify a candidate dynamic case assignment scheme aimed at effectively countering the most deleterious effects of workload variability.

1.2 Project motivation, overview, and key insights

From a high-level, ostensibly objective, perspective the genesis of this project can be accurately portrayed as a variant of a ubiquitous resource allocation problem formulated as:

> *Given* a finite number of available weekly case manager hours to allocate between two main case manager functions and across 43 hospital floors with different patient populations,
>
> *What* is the proper way to allocate case manager hours to achieve an equitable division of workload among case managers in a way that limits the deleterious effects of inevitable workload variability?

While the succinct, encompassing objectivity of the above formulation cannot be faulted, this may be an overly abstract way to frame a problem influenced greatly by interpersonal interactions, patient psycho-social characteristics, family dynamics, and other factors that do not easily yield to quantification. In fact, despite the necessity of abstractions and abstract analysis to make any progress on the problem at hand, from the frontline case manager perspective it is just as accurate to portray the content of this thesis as stemming from attempts to "solve" a morale problem.

To briefly explain, case managers, as described in the previous section (and in greater detail in Chapter 2) perform one of two primary functions: utilization review and discharge planning. Within the discharge planning function different case managers are responsible for cases with varying case characteristics; the "typical" case on one floor may take more or less work than the "typical" case on another floor. The potentially unequal distribution of work may plant the seeds for a problem.

The preceding should in no way be interpreted as an indictment of case managers because, based on extensive observation of case managers, even if work may not equitably distributed, this in no way affects faithful and tireless execution of their mission to ensure a patient's care needs are met, both while at MGH and upon discharge. Often this work ethic is so pronounced that it can make it difficult to discern a case manager with a caseload that is excessive; the mission is still completed. It was in recognition of this work ethic, whether directly observed or inferred from reading thousands of case notes, which leads to a third way to state the guiding principle of this thesis – "What can our team do to improve the daily work conditions faced by inpatient case managers at MGH?"

No matter how the problem is formulated, the fundamental requirement for progress toward a solution is an ability to meaningfully measure case manager workload. This ability to measure workload allows the requisite current state analysis. Furthermore, a measure of work, referred to as a work score or work metric in this thesis, can serve as a response variable for explanatory or predictive modeling at the case or day level. Figure 1-1 provides on overview of the major phases and sub-phases of our work, although the figure implies a strictly linear course that was not always the norm.

The initial tack we employed was simple, though admittedly tedious and time-consuming. More than three thousand case manager progress notes for 1500+ cases were reviewed to identify work events documented in the notes. Examples of these work events include, but are not limited to, phone calls to various entities, meetings with patients and families, and referrals placed to post-MGH care providers. The initial scope of the project was limited; two general medicine floors were investigated to develop the techniques necessitated by our aim. A simple scoring system, based on observation and interviews with case managers, was developed to allow assignment of a work score to each progress note, each case (which may contain multiple notes), and each day (which contains notes from multiple cases).

The work score serves a foundational purpose for the analysis presented in this thesis. Given this importance, the assistance of two experienced case managers, unaffiliated with the cases examined, was enlisted to help validate the scoring methodology. Each case manager was provided 20 cases for review across the spectrum of scores. The case managers ranked the cases in order of increasing workload and these ordinal rankings were compared to the ordinal rankings derived from our work score. Comparing the rankings yielded a Kendall's W (Kendall's coefficient of concordance) of 0.98, indicating substantial agreement between the work score we developed and case manager rankings, at least on an ordinal scale. Further refinement and validation is necessary to fully calibrate the

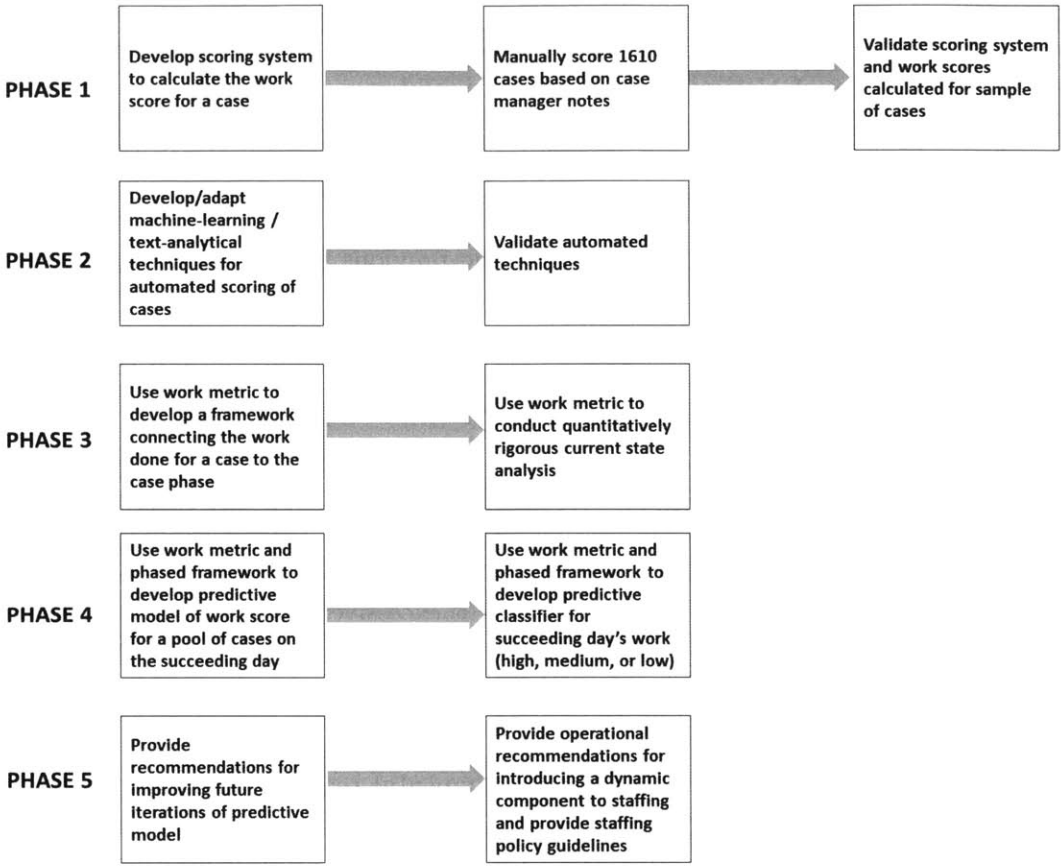


Figure 1-1: Overview of research phases and sub-phases

work metric scale in units of time, but the current validation procedure provided strong support for the utility of the current scoring system.

Further support for the aptness of the current scoring methodology emerges when the distribution of work between cases and throughout the workweek is examined. As reported by case managers a relatively small number of cases consumed the greatest amount of case manager time. The top decile of patients, by our work score, accounted for 40% of the total work scored. In contrast, the bottom 60% of patients by work score accounted for on the order of 20% of this total, while the 61^{st}-90^{th} percentile cases accounted for the remaining 40% of this total. The overall weekly periodicity was also evident when using our work score, with Mondays and Fridays exhibiting a marked tendency to be the highest workload days. This periodicity is easily explained by the differential staffing patterns on the weekend vis-à-vis weekdays as explained in this thesis.

The time-consuming nature of manually scoring notes and cases imposes significant limitations on extending scoring to other floors. Therefore, the next step in our work was developing a text-analytical engine to allow reliable automated scoring of case manager notes. The heart of this engine relies on an augmented bag-of-words (BOW) approach. As described in detail in Chapter 4, a basic textual analysis was completed to identify words indicative of a workload event. This allowed us to construct a dictionary to use in our analytical engine. This super-dictionary was refined to form dictionaries for different word categories. A count of the words in these constituent categories formed fifteen fields of a 29 feature vector constructed for each case note. Other key fields

of this vector included an overall dictionary word count and a header for the type of note. Together, these fields provided the augmentation for our BOW-based analytical engine; the mechanism of this augmentation is in preserving the relevant context for each note, a context that may be lost with a traditional BOW approach. Using a regression tree we were able to automatically score a validation set of cases, over multiple trials with a 60/40 training/validation split, with an R^2 in excess of 0.97. Similar performance was achieved at the day-level.

Armed with a validated work score, the next phase of our work focuses on a current state analysis for the two floors examined. Several key insights emerge from this analysis. First, it is possible to identify distinct phases that a case progresses through on the discharge planning plane, including an admit window, a pre-discharge window, and a discharge window. The state of each case within these phases can be specified at an even more granular level. The phased framework we develop greatly facilitates the current state analysis and our predictive modeling efforts, introduced below. The analysis, described in Chapter 5, also indicates that the top decile of cases by work score, though often greatly outstripping other case groups in terms of aggregate work, appear remarkably similar to other cases in terms of work done on any given day. When this observation was paired with the fact only a subset of cases are active on any given day, our concept of an active census emerges. This concept is a cornerstone of our predictive modeling.

There are many sources of day-to-day variability in case manager workload. The obvious source of variability arises at the case-level because of differences in case characteristics, including needs upon discharge, myriad psycho-social factors, and patient insurance. There is also variability between floors that can be traced to extra-CM work patterns and processes; the case manager work processes are embedded in a larger system of floor processes so that it is possible a case could require more work on one floor than on another floor. However, the greatest source of variability is whether a case is active on any given day or not; that is, does a case manager perform work for a specific case on a given day? This touches on the key insight from our predictive modeling – the timing of work (is a case active / how many cases are active?) and the general nature of the work (is the work performed in the admit window or the discharge window?) explains most of the variability observed in a case manager's daily workload. Using a conventional interpretation of the coefficient of determination, R^2, these factors account for 66% of the observed daily workload variability.

This is included in our linear-regression based modeling by using a count of the number of patients in each state/phase at 04:00 to predict the workload for the upcoming day. On validation sets admit window work can be predicted with an adjusted R^2 of 0.66 and total work with an adjusted R^2 of 0.51. One key patient state/phase used was a count of discharge window cases active the previous day. The tendency of these cases to be active the current day was an exploitable pattern for modeling purposes. In essence, this introduces an auto-regressive component into modeling, a component that exists because of delays characteristic of a request-response cycle. When developing and executing a discharge plan case managers often query or coordinate with (request) potential post-MGH care providers / sub-acute facilities. As the post-MGH entities review the case a delay is inevitable before a response is provided to the case manager. Based on this response the case manager continues with the current plan or alters the plan.

Other patterns prove not to be currently exploitable for predictive modeling purposes. Further segmentation of cases into, for example, an early and late pre-discharge window, or by incorporating outcomes of the previous day's work for a case, will enhance the capability of a regression-based model. Potentially promising avenues of inquiry, based on preliminary investigation, are explicitly

identified in Chapter 6. The first key to improving the model is in improving the ability to predict when a case will be active. This explains why even knowing with perfect certainty that a case will require a large total amount of work over the course of a patient's entire length of stay is a very poor predictor of workload on any specific day. Thus, a count of the high-workload census for a case manager is unsuitable as a reliable signal to "flex" any available resources to a particular floor.

In addition to a regression-based model, a classifier based on a boosted classification tree was examined. This classification problem may ultimately prove to be more tractable than a model designed to provide a predicted workload score for the current day. An intricate procedure, employing hierarchal clustering for partitioning the data set into training and validation sets, synthetic minority class oversampling (SMOTE), one-sided selection, condensed nearest neighbors, and Tomek-link concepts was used to train the boosted classification tree[54][87][125][77][3]. The results were promising, but the intricacy of the training procedure may be biased to perform well on the data set examined. Still, on validation data sets with a high, medium, and low workload classification scheme only 6.9% of days exhibited a two-class misclassification (high-as-low or low-as-high). The overall classification accuracy using the three-level classifier exceeded 81%.

The final phase of our work includes preliminary efforts at both developing staffing schemes to equitably distribute workload across floors and case managers, and introducing a dynamic component to staffing. Considering the former, a procedure to form a CM-specific index analogous to case mix indices based on clinical factors[26], and related to hospital resource utilization is provided in Appendix B. In effect, this index would allow a "typical" case on one type of floor to be compared to a "typical" case on a different type of floor[4]. By facilitating cross-floor comparisons this would allow appropriate relative benchmark caseloads to be established across floors.

Considering dynamic elements, pooling is examined, not only as a means to decrease total workload variability, but also, and primarily, to facilitate a mechanism by which the magnitude of work, on high-workload days, could be attenuated. This work is very preliminary, but the effects of pooling on decreasing variability are marginal or equivocal when looking at traditional measures. In fact, traditional measures, such as standard deviation or measures derived using standard deviation, may not reflect how case managers experience variability in workload. This is true of any measure that treats upside and downside workload variability symmetrically. Still, pooling floors, treating them as a unit with multiple case managers available for coverage, offers potential benefits for which a limit of performance can be specified. Drawing on all phases of this work a "one-switch" dynamic case assignment scheme (1SDCA) is outlined. Here, the key is to balance, as nearly as possible, the number of cases in each phase that pooled case managers are responsible for. Cases are initially assigned with the aim of balancing unassessed (newly admitted) patients. This is the tentative case assignment. A "one-switch" is possible when a case is ready to enter the discharge window to balance the number of discharge window cases between pooled case managers. Limiting the reassignment

[3]In fact, many other techniques were employed over the course of of our work in an attempt to deal effectively with imbalanced data sets. Other techniques are discussed in Chapters 3 and 6, as well as Appendix A. In many of our sub-problem formulations for classification there is a clear majority class and minority class; e.g., 10% high workload days and 90% low workload days. The work of Chawla, as well as the work of He and Ma, provides a solid introduction to the problems associated with imbalanced data sets[53][78]. Branco et al. provide a very accessible introduction to predictive modeling under conditions of distribution imbalance[43]. The primary discussion of potential issues from the practical perspective of real-world costs associated with misclassification is taken up in Chapter 6.

[4]Typical, in the sense of average, only has utility as a concept over long time horizons. As discussed in Chapters 5 and 6, even long time horizons afford little value to the concept of a typical case for the purposes of predicting daily workload.

(switch) to, at most, one episode per case, at the beginning of the discharge window[5] mitigates most of the undesirable impacts associated with the loss of case manager-patient continuity.

1.3 Thesis organization and structure

Chapter 2 expands on Section 1.1.3 by describing in greater detail case management at MGH. This chapter is important in detailing the context in which our work was performed. The practice of case management is not standardized across institutions so it becomes vital to define inpatient case management as practiced at MGH in a tertiary hospital setting. The thesis continues with a literature review in Chapter 3, beginning with a review of systems proposed for quantifying the amount of work required of a case manager for a case with a given set of characteristics. These systems are almost exclusively formulated in terms of an acuity score. Chapter 3 also reviews recent efforts at employing text-analytical techniques on hospital documentation. These efforts are not specific to the realm of case management, but the review is instructive given the importance of our text-analytical engine for automatically scoring cases. Finally, Chapter 3 provides a review of key machine learning techniques used in our work; this part of the review is limited to techniques that, in our estimation, may be less familiar to readers.

Chapter 4 is the foundational chapter for this thesis. Here a detailed description of how we constructed the work score is provided, as is a similarly detailed description of our validation procedure and results. The chapter then presents our text-analytical engine used for automated scoring as well as the performance of our automated scoring methodology. Chapter 5 provides a current state analysis using the validated work metric. This chapter is key in introducing the concept of an active census, as well as the framework developed to unambiguously determine which phase a case is in; if Chapter 4 is viewed as the foundation of this work, then Chapter 5 is the bridge that facilitates predictive modeling. Chapter 6 employs the phased framework, specifically to obtain counts of how many cases are in each phase, to develop a regression-based predictive model of case manager workload for the upcoming day. As an alternative to this regression-based model, the performance of a classifier based on a boosted classification tree is examined.

The thesis concludes in Chapter 7 with recommendations, both for future work as well as operational recommendations. In truth, recommendations are not confined to Chapter 7. Because a large portion of our work revolved around developing techniques specific to our problem domain, it is more efficacious to present recommendations in context; i.e., immediately following the discussion where the recommendations naturally arise. Chapter 7 recounts the most important of the recommendations for improving both an ability to measure and an ability to predict CM workload. The operational recommendations are preliminary, but the CM-specific case-mix index in Appendix B, as well as the 1SDCA scheme outlined in Chapter 7, may be the most promising in the near-term.

1.4 Potential methodological extensions of thesis

Though our work was conducted in a very specific setting, inpatient case management at MGH, it may be possible to adapt some aspects of our approach to other settings. In fact, quantifying

[5]Our discharge window convention is introduced in Chapter 2 and rigorously defined in Chapter 5.

workload as a preliminary step to predicting workload and making better informed staffing decisions has applicability in a wide-range of service industry settings. Time-motion studies and methods engineering approaches, of course, are already used in service organizations such as hospitals. Yet, if the organization already possesses a large amount of data, including unstructured free-text data, created during the normal course of business, then the techniques we describe may allow an organization to leverage the use of this data.

While we cannot specify exactly how our approach should be modified for another setting, it is possible to identify setting characteristics that may facilitate technique adaptation. The concept of a case has relevance for many realms - law enforcement/investigation, social work, IT help desks, legal work, and insurance adjustment, to name a few. One characteristic of an environment where our techniques may adapt well is discontiguous work. In discontiguous work a case is not worked from start to finish. Rather, work proceeds in fits and starts, either because the information discovery process allowing subsequent work to proceed, or a request-response cycle, introduces delays.

Another feature suitable for technique adaptation would be when caseload does not correlate strongly with workload. In some ways this is another result of discontiguous work. More specifically, on a given day there could be more or fewer cases for which work is required or possible. Work environments where distinct phases can be identified also offer the possibility of technique adaptation, particularly when predicting workload. The phases, requiring the development of definitions and conventions, can also help provide structure for unstructured data. In the context of our work certain phases we defined were associated with either more work for a case or a higher probability that a case would require work on a given day.

As service industries and organizations become a larger part of the economy, effectively quantifying workload as done in our work assumes ever greater importance. Quantifying workload when "non-routine" service industry work tasks are considered presents challenges as compared to, for example, manufacturing settings. Here we are using non-routine in a manner similar to Autor et al. to refer to tasks requiring "problem-solving, creativity, intuition, and interpersonal skills[35][34]." Though non-routine these skills are often exercised in the context of relativey few work events such as phone calls, meetings, emails, and documentation.

From case to case the frequency, duration, composition, and timing of these work events can vary depending on specific case features and complex interactions among these features. As we show, it is possible to extract information from data available in a service organization to quantify the work done for a case despite inter-case variability. This quantification is then available to facilitate predictive modeling and inform staffing decisions to achieve an equal distribution of work and address workload variability.

Chapter 2

Inpatient Case Management at MGH

Case management at MGH centers on the performance of two primary tasks: utilization review (UR) and discharge planning (DCP). Both of these tasks are critically important in helping to control hospital costs and maximize revenue while, more importantly, ensuring that the quality of patient care is not compromised. At the most basic level UR is focused on ensuring efficient, effective, and appropriately reimbursed use of MGH's limited capacity and care resources. DCP's focus is on ensuring safe and timely transitions of patients discharging from MGH to post-acute care settings that meet patient needs; these settings comprise a wide-variety of facility types and home-care plans. The DCP and UR roles are mutually reinforcing.

It is easy to place and appreciate the importance of UR and DCP in the context of a number of general trends in American healthcare. One well-documented, broad trend over the past four-plus decades is the marked decrease in the average hospital length-of-stay (LOS), particularly for older patients. As an example, among individuals 65 years of age and older, the average hospital LOS across all conditions has decreased from 12.6 days in 1970 to 5.5 days in 2010[70].

There are many drivers for this reduction in LOS, including the advent of Medicare's prospective payment system and the growth in post-acute / subacute care options and facilities[75]. DCP and UR have likewise played a role in this broader trend, as well as becoming increasingly necessary as the trend progressed[75]. For example, as stated, DCP case managers help ensure the safe transition of patients to the more appropriate and cost-effective post-acute setting that can meet a patient's post-MGH care needs that, though not requiring acute care, may still be complex. Safe and timely transitions help improve community access to acute care at MGH.

At MGH utilization review and discharge planning are performed by a dedicated team of registered nurses serving as case managers. In their functional UR and DCP roles, case managers also assume other more general roles as assessor, planner, facilitator, and advocate[11]. No matter the role, titular or more general, case managers have to constantly strive to balance individual patient needs, preferences, and insurance considerations, all while supporting the broader MGH priorities and mission.

This chapter begins with a consideration of the current MGH case management model, placing this model in the typology of models discussed in the literature[49][50][51][52][95]. UR and DCP are then, in turn, discussed in greater detail with a focus on how these functions are performed at

MGH. This includes a discussion of key interfaces with other entities and functions both internal and external to MGH. Following this discussion the CM division of labor and resource deployment, both between DCP and UR and within each of these sub-functions, is presented.

The focus of the chapter is then narrowed to consider DCP exclusively as this was the primary domain of our research. Potential sources of variability in an individual discharge planner's daily workload, as well as workload variability across hospital floors, over an extended time horizon, are identified. The penultimate section of this chapter provides a brief description of data sources used in our work. Finally, explicit identification of our problem statement is provided, a problem statement informed by our belief that a case manager's time and accumulated experience are the most valuable of CM resources.

2.1 MGH case management model

As noted in the literature, there is no standard model of case management in an acute care setting such as MGH[14][49][95]. In fact, as of 2013 there were no fewer than 6 different accrediting agencies for case managers and 21 case management-related certifications[95][90][126]. There are numerous terms to describe case management models (e.g., integrated case management, partially-integrated case management, embedded case management, etc.) but Cesta has suggested that there are three primary models of case management in an acute setting like MGH: partially integrated dyad model, integrated model, and collaborative triad model[49].

The traditional purview of case management consisted of two domains, utilization review and discharge planning, domains which are the current focus for case management at MGH. Note that Figure 2-1 through 2-3 are adapted from Cesta and Mawn[49][95].

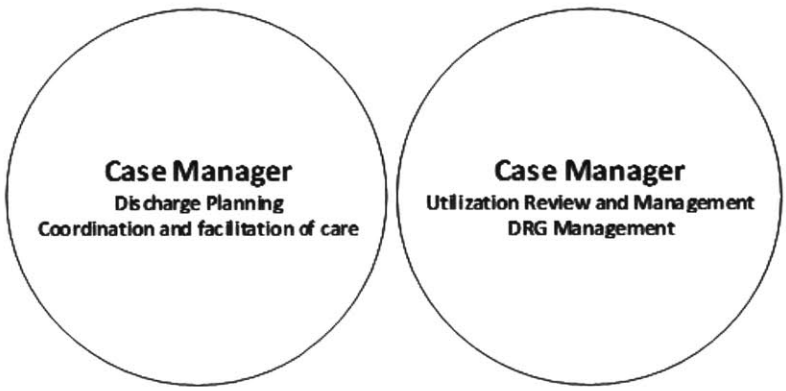


Figure 2-1: Traditional silos of case management - *adapted from [49][95]*

Before reorganization began in June 2014 the MGH case management model followed a variant of the traditional dyad model, with a single case manager performing both DCP and UR activities for a patient and other members of the care team supporting discharge planning. The previous model is identified as a variant of the canonical dyad model because, typically, the other part of the dyad is envisioned as a social worker that attends to the psychosocial aspects of discharge planning. The other part of the dyad really depends on the patient. Furthermore, based on our observations at MGH it is not usually accurate, nor advantageous, to segregate discharge planning into psychosocial

domains and "other" domains as portrayed in Figure 2-1. It is noteworthy that the only exception to this ill-advised segregation of DCP activities is on Blake 11, the inpatient psychiatric floor at MGH. For Blake 11 patients the psychosocial aspect and UR/DCP activities may be more clearly separable and, despite the CM reorganization throughout the rest of MGH, Blake 11 most clearly adheres to a traditional dyad model with one CM handling discharge planning and UR for 24 patients, while three social workers attend to the psychosocial aspects of DCP, including family meetings, for eight patients each.

Figure 2-2: Traditional dyad model of case management - *adapted from [49][95]*

Beginning in 2014 MGH moved to a model of case management in which UR activities and discharge planning were separated among different groups of case managers. This model of case management is a variant of the triad model identified in the literature. Again, the modifier variant is used because the third part of the triad may change depending on the patient (case) and it is usually difficult to separate out the psychosocial aspects of discharge planning.

It is important to point out that the current model of case management at MGH is referred to as a dyad model with the dyad composed of UR case managers and DCP case managers. The designation is at odds with some of the literature but, in this case the difference is purely semantic. The semantic difference does hint at a potentially salient aspect of MGH organization. To explain, the registered nurses acting as case managers ultimately have a different reporting structure than other registered nurses and social workers.

The shift to the current MGH dyad model was prompted, in part, by the positive experiences of other hospitals using the dyad model, such as Cleveland Clinic and Baystate Medical Center[8]. With the shift DCP case managers had more patients but did not have to perform UR activities. UR case managers had more cases (typically twice as many as DCP case managers) but performed only UR activities. The reaction of payors to the new model has been positive and there have been fewer denials[8][6]. However, interviews with DCP case managers indicate the trade-offs with the new model as they typically report experiencing a higher workload and the need to be reactive rather than having opportunities to be proactive in attending to cases[7][1][1]. The rollout of the dyad model occurred in a stepwise fashion beginning on 8 June 2014. During the period covered by this thesis, the rollout was nearly complete with the exception of Blake 11, as previously mentioned, the obstetrics floors on Blake 13 and Ellison 13, and the Blake 14 labor and delivery floor. One case manager handles UR and aspects of DCP on Blake 11 while another case manager handles UR and

[1]This sentiment was also expressed by many of the other case managers interviewed over the course of this work; however, express permission was not obtained to allow attribution in this thesis.

Figure 2-3: Traditional triad (collaborative) model of case management - *adapted from [49][95]*

DCP for Blake 13, Blake 14, and Ellison 13.

Regardless of the terminology used, the current MGH dyad model of case management is more accurately represented by Figure 2-4 from a DCP case manager's perspective. Conceptually, the degree of overlap (area of intersection) with the DCP case manager depends on both the floor and the patient. Similarly, the size of each actor's circle depends on the patient.

2.2 Utilization review overview

As explained in the chapter introduction, the focus of utilization review (also commonly referred to as utilization management) is ensuring efficient, effective, and appropriately reimbursed use of care resources. Depending on the setting, UR may have prospective, retrospective and concurrent aspects[25]. At MGH formal UR focuses on retrospective and concurrent reviews.

Retrospective reviews are typically conducted within 24 hours of patient admission to the hospital and are designed to ensure that a patient was admitted correctly, either as "observation" or "inpatient"[18][22][24]. For patients with insurance requiring an authorization process these initial reviews help confirm that the physician's admitting order and payor authorization for patient status and level of care (LOC) match. Concurrent reviews are performed during a patient's stay to ensure that a patient's stay remains a medical necessity and that the level of care being delivered warrants continued hospitalization. The concurrent clinical reviews are provided to payors at a frequency as specified by contract throughout a patient's stay.

UR documentation is used by payors to authorize reimbursement and, as registered nurses, UR case managers combine clinical knowledge with knowledge of reimbursement mechanisms, provider

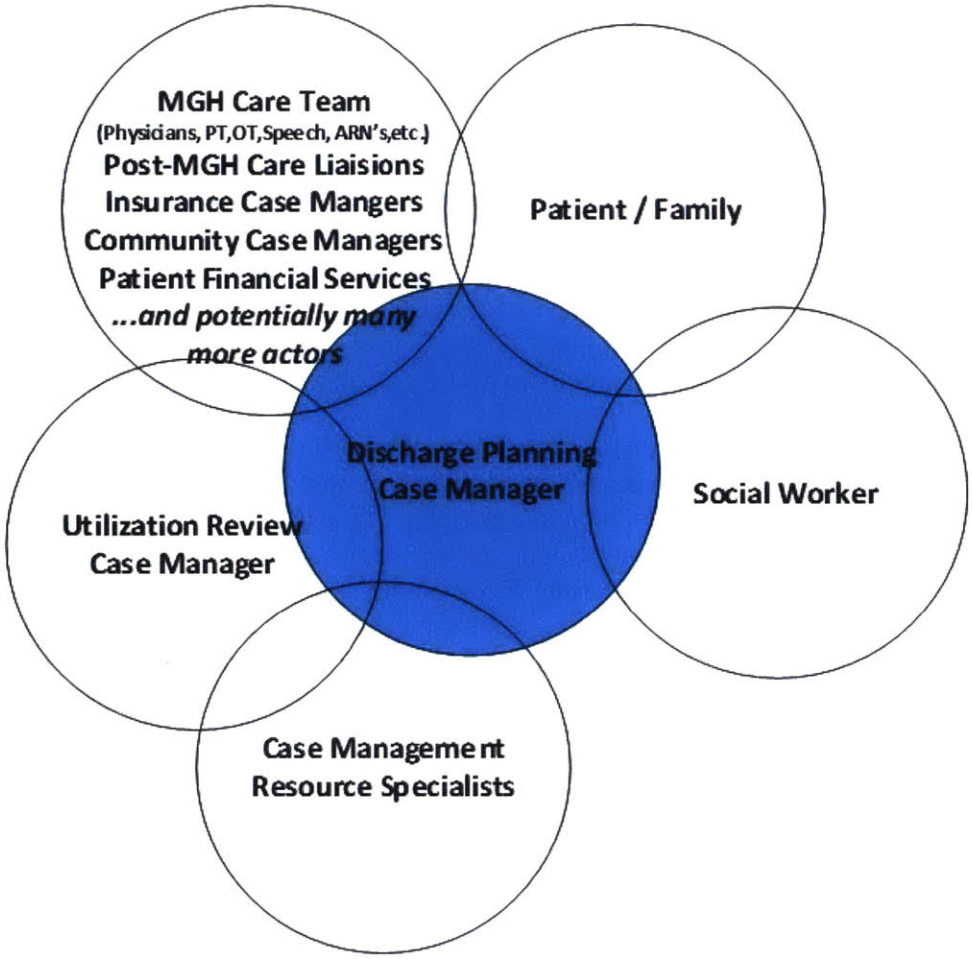


Figure 2-4: Conceptual representation of case management at MGH from a discharge planning perspective

options, payor benefits, and payor requirements to create an effective skillset with benefits for the hospital, patients, and payors. From a high-level perspective UR provides a component of the mechanism to ensure that MGH's limited capacity and resources are used to meet the needs of patients requiring acute care, rather than patients whose needs can be met by sub-acute facilities. Figure 2-5 shows a generalized process map for UR at MGH based on CM interviews and observation of UR case managers.

Referencing the process map, and assuming a non-elective, unplanned, patient presentation to the hospital for expository conciseness (e.g., a patient first seen in the emergency department (ED)), the movement of the patient can be along one of three primary paths. The patient may be treated and discharged from the ED. Alternately, the patient's condition and/or required interventions, such as diagnostic tests, may warrant a LOS longer than a typical ED encounter. The patient may be admitted as an observation case. Finally, a patient's condition may require an inpatient admission if the attending physician believes a patient's condition will require a stay exceeding some threshold. Payors also have varying thresholds dictating when a patient initially admitted for observation should be subsequently admitted as an inpatient. From an outside perspective this threshold varies by payor (e.g. 48 hours for Medicare with decision to admit as inpatient typically within 24 hours,

48 hours for Medicaid, variable for private insurance but usually within 23 hours)[20][18]. From MGH's perspective if the observation period exceeds 24 hours then an inpatient admission would usually be triggered.

Though decision-making aids exist to help determine whether a patient should be admitted for observation or as an inpatient, the initial decision during a patient's stay is not always clear. From both a payor and hospital perspective the decision-making authority resides with the admitting physician[24]. The decision to admit as observation or inpatient is not merely semantic. That is, though the quality of care received is the same under either designation, the admission categorization has consequences for hospital reimbursement, patient out-of-pocket expenses, hospital regulatory compliance, and the time/expense associated with any re-categorizations[24][18].

Though the rules are complex and vary by payor some examples of categorization consequences can be considered. If observation services are ordered the hospital would receive much less reimbursement than for equivalent services rendered for an inpatient[24]. The patient, receiving outpatient services, would also be responsible for any co-insurance payments for outpatient services, payments that may exceed their inpatient deductible[18]. However, if, for example, a Medicare patient is admitted when they should have received outpatient services then Medicare auditors could deny the claim in its entirety, causing the hospital to lose reimbursement and incur the additional expense of appealing the denial[24]. Furthermore, the time a Medicare patient spends in observation does not count toward the "three consecutive midnight rule" to qualify for a covered skilled nursing facility (SNF) stay, and a patient requiring a lower level of care such as SNF placement could bear responsibility for the costs[20][28]. Finally, auditors tend to focus on short hospital stays (one, two, or three days) to find inpatient admissions that may be denied as unnecessary. A high frequency of short inpatient admissions could trigger scrutiny and the sanctions for admission categorization non-compliance could be severe[24].

The preceding discussion is meant to provide some appreciation for the importance of and complexities surrounding UR. Returning to the process map, the initial retrospective review occurs within 24 business hours. McKesson's InterQual criteria are used to help verify categorization, medical necessity and LOC[99][96][19]. The initial review information is then provided, as required, to payors according to any authorization process to demonstrate a patient meets medical necessity and warrants an acute LOC. The range of a UR case manager's skill set, from clinical knowledge to knowledge of specific payor processes, is brought to bear during initial review of a new case in preparing documentation that will be examined by a similarly trained counterpart in employ of the payor. Re-categorization from inpatient to observation, or vice versa may occur. In the context of Medicare, the single largest payor at MGH, the former re-categorization requires meeting very strict criteria and is expected to occur infrequently; the latter re-categorization can occur at any time[24]. Again, the process for initial review can vary slightly between payors, both public versus private and across the range of private payors.

Concurrent, or ongoing, reviews are conducted periodically to ensure that a patient continues to meet medical necessity and LOC criteria. The interval of review may be payor-dependent. InterQual is still available for use as a decision-support tool, but, in practice at MGH, the clinical review documentation completed using the UR case manager's knowledge and expertise is more important than InterQual for concurrent reviews[4][2]. The InterQual evaluation may still be completed for

[2]As the focus of our work was discharge planning, observations of case managers performing UR activites were comparatively limited.

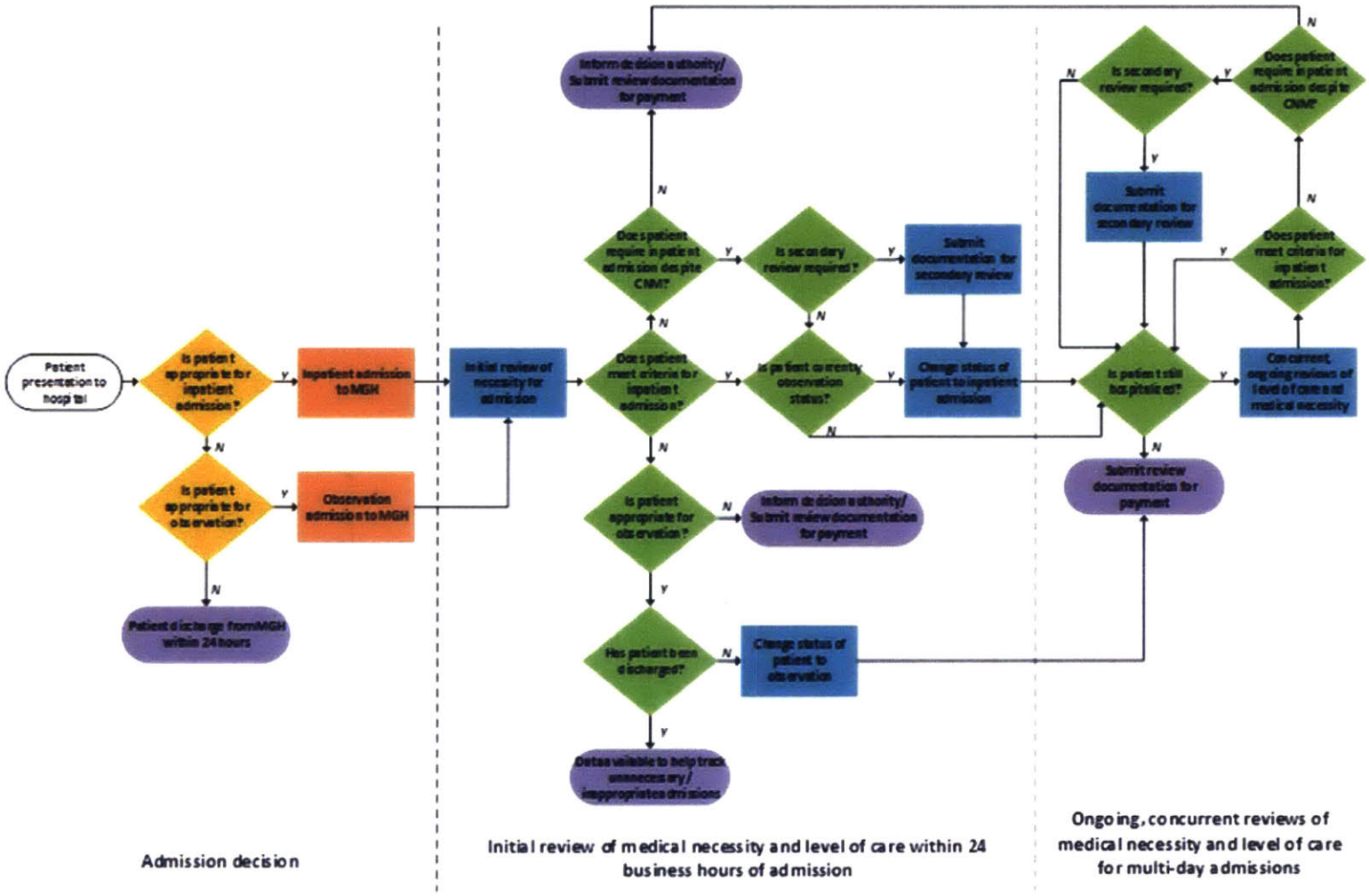


Figure 2-5: Generalized utilization review process map

the review if the case for continuing hospitalization, based on clinical information, is not clear-cut. However, this InterQual evaluation is in support of the clinical review and seems to be included only because some payors may put greater weight on an ostensibly objective InterQual evaluation in the case of any ambiguity.

In a similar vein, the judgement of the UR case manager assumes paramount importance when criteria for continued hospitalization are not met but, in the judgement of the care team, the patient still needs to remain hospitalized. This can occur if there are concerns about a patient's safety post-discharge. Such cases often necessitate a secondary review as indicated in Figure 2.1. The UR case manager, synthesizing information from clinical documentation and communication from the care team, provides the synopsis for these secondary reviews. In other cases a patient may not meet criteria for continued hospitalization, but a secondary review is not required. For example, a patient may be ready for discharge to a post-acute facility but there is no facility to accept the patient. In terms of InterQual criteria this would be coded as CNM-SD ("Criteria Not Met – System Delay") and a secondary review may not be required[4].

The UR case manager spends a majority of time synthesizing information from a variety of clinical documentation, acting to protect the interests of the hospital in terms of reimbursement and compliance and, importantly, acting as an unseen advocate for the patient to ensure they receive needed care, covered under insurance, by justifying the medical necessity of a stay to payors. Our observations of UR at MGH were limited, but it is clear that in the UR role the clinical knowledge that RN case managers possess is indispensable. The UR case managers observed tended to prioritize their work beginning with initial retrospective review of new cases, proceeding to review of Observation cases, and then concurrent reviews. The amount of time spent on each activity could vary based on the day of the week as a consequence of limited weekend staffing.

UR case managers interact with the care team as needed, as well as with Medicare leveler teams at MGH and Case Management Resource Specialists (CMRS). Given the predominance of Medicare as a payor at MGH the leveler teams, composed of registered nurses at MGH who review all Medicare cases, are a valuable resource for the UR case manager. CMRS personnel play a role in mediating the exchange of information between UR case managers and payors. This includes faxing clinical review information to payors, faxing daily census and discharge reports to payors, manning a UR line that payors can call to request additional information on a patient, and helping review UR documentation for billing purposes[2].

2.3 Discharge Planning Overview

As stated in the introduction of this chapter, the unrelenting pressure to reduce healthcare costs while maintaining the quality of patient care has led to drastic reductions in hospital lengths-of-stay[75]. True, advances in medical technology, and less invasive medical and surgical techniques, has helped obviate the need for a portion of some patients' LOS, but in many cases a patient's care needs extend past the point of discharge from an acute care setting like MGH. In order to meet continuing patient needs, discharge planners, in consultation with the care team, family, payor representatives, and potentially many more actors, develop discharge plans that ensure a safe transition to a sub-acute level of care, including home plans. What is more, discharge planning for MGH is not discretionary. Instead, it is a condition of participation if receiving reimbursement from Medicare or Medicaid

programs[10][69].

While there may be common themes among discharge plans for patients, each plan is tailored specifically for a patient. The variation in patient needs and non-routine aspects of discharge planning introduces a certain complexity to DCP that may, in fact, be uncorrelated with a patient's clinical complexity. From a high-level perspective this complexity can be traced to a discharge planner's requirement to balance patient needs, preferences, insurance/payor considerations, and post-MGH provider preferences, all in the context of hospital priorities. For purposes of illustration this balancing is illustrated sequentially in Figure 2-6.

As long as all of the factors remain aligned then discharge planning can usually proceed along an uninterrupted path to completion. Certainly, there are some other system-wide considerations that can affect a patient's path to discharge, particularly in the case of a limited capacity of long-term acute care (LTAC) or inpatient psychiatric facility beds, and introduce delays and/or rework. However, in general, alignment means proceeding through each decision point only once.

Figure 2-6 shows only one explicit sub-process, the case manager meeting with the patient, but each decision point also implies work on the part of the discharge planner. That is, it takes time in the form of communication and coordination with other entities to gather the information required at each decision point. When relevant considerations become misaligned this can require the case manager to restart the DCP process. One other element implicit in Figure 2-6 is time. To explain, a patient's needs and, in some cases even a patient's preferences or insurance, can change over the course of the patient's stay. These dynamic, evolving conditions can also cause the process of factor alignment to restart. What is more, a discharge planner may have 20+ cases, each at a different point in Figure 2-6.

Of course, Figure 2-6 is a very high-level overview of the DCP process and, as such, the complexity of discharge planning is grossly understated. The primary reason for this is an inadequate representation of the patient needs concept. To be sure, a patient has clinical needs, but psychosocial factors are usually the determining factor in how complex and how much time a discharge plan takes to develop and execute[14][38][91]. Figure 2-7, adapted from a Case Management Society of America working paper affords some appreciation of how encompassing the concept of patient needs is[14]. It bears noting that interviews with case managers, in addition to the cited literature, suggest that these psychosocial factors are much more salient considerations in determining how much time it takes for discharge planning[6][7]. Not all of the factors shown in Figure 2-7 are relevant considerations for each patient. However, the important point to bear in mind is that each factor must potentially be considered in developing a discharge plan. In fact each factor was relevant in at least one of the 1600+ cases that we examined during the course of this work.

While an understanding and appreciation for Figure 2-7 provides crucial insights into the day-to-day complexities faced by DCP case managers at MGH, the high-level perspective admittedly precludes practical insights. Figure 2-8 illustrates patient flow from the DCP case management perspective. It is important to note that patient progression can be traced along three separate planes: clinical treatment plane, discharge planning plane, utilization review plane. The DCP case manager at MGH provides the necessary linkage between these three planes.

Referencing Figure 2-8, when there is a new inpatient admission the patient is initially an unassessed patient. It is possible for an inpatient to discharge from this state and this usually occurs if a patient is admitted on the weekend (beginning Friday after approximately 1730) and discharges

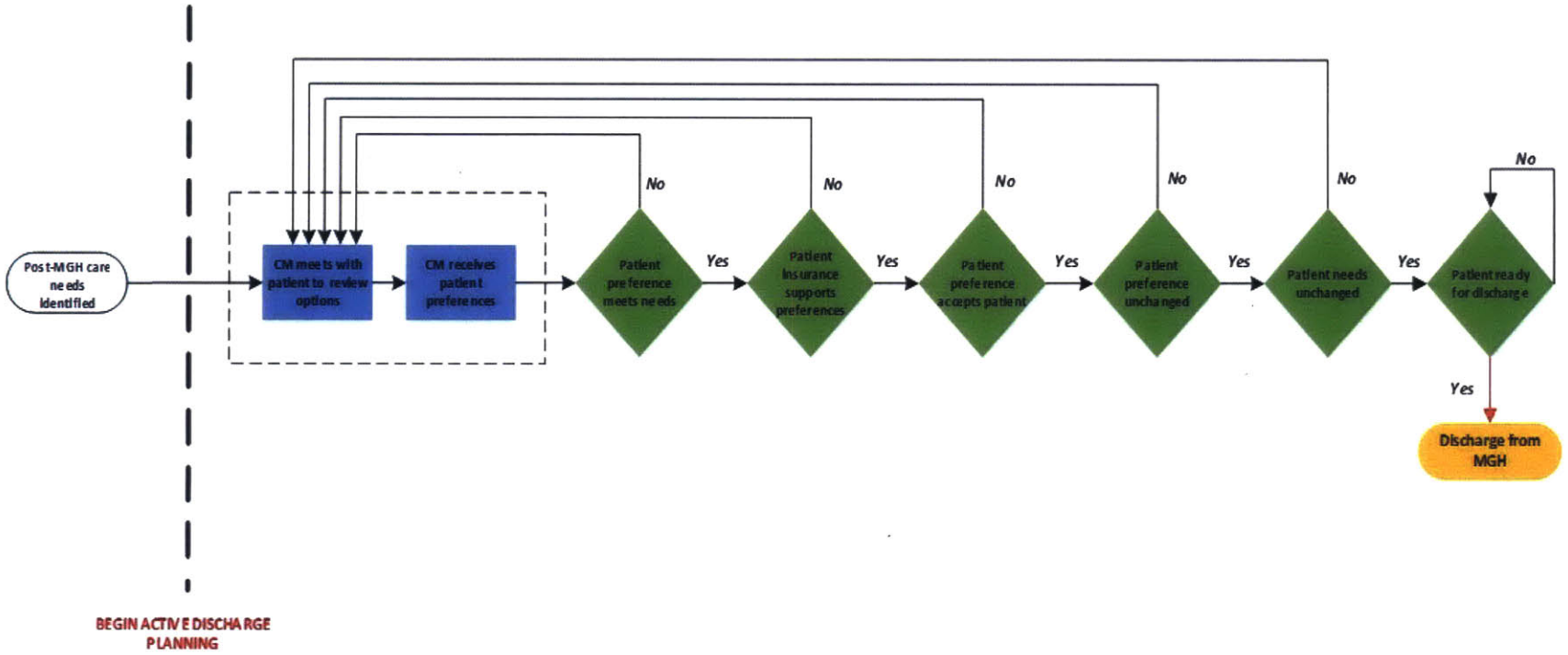


Figure 2-6: Aligning factors to facilitate a safe and acceptable discharge

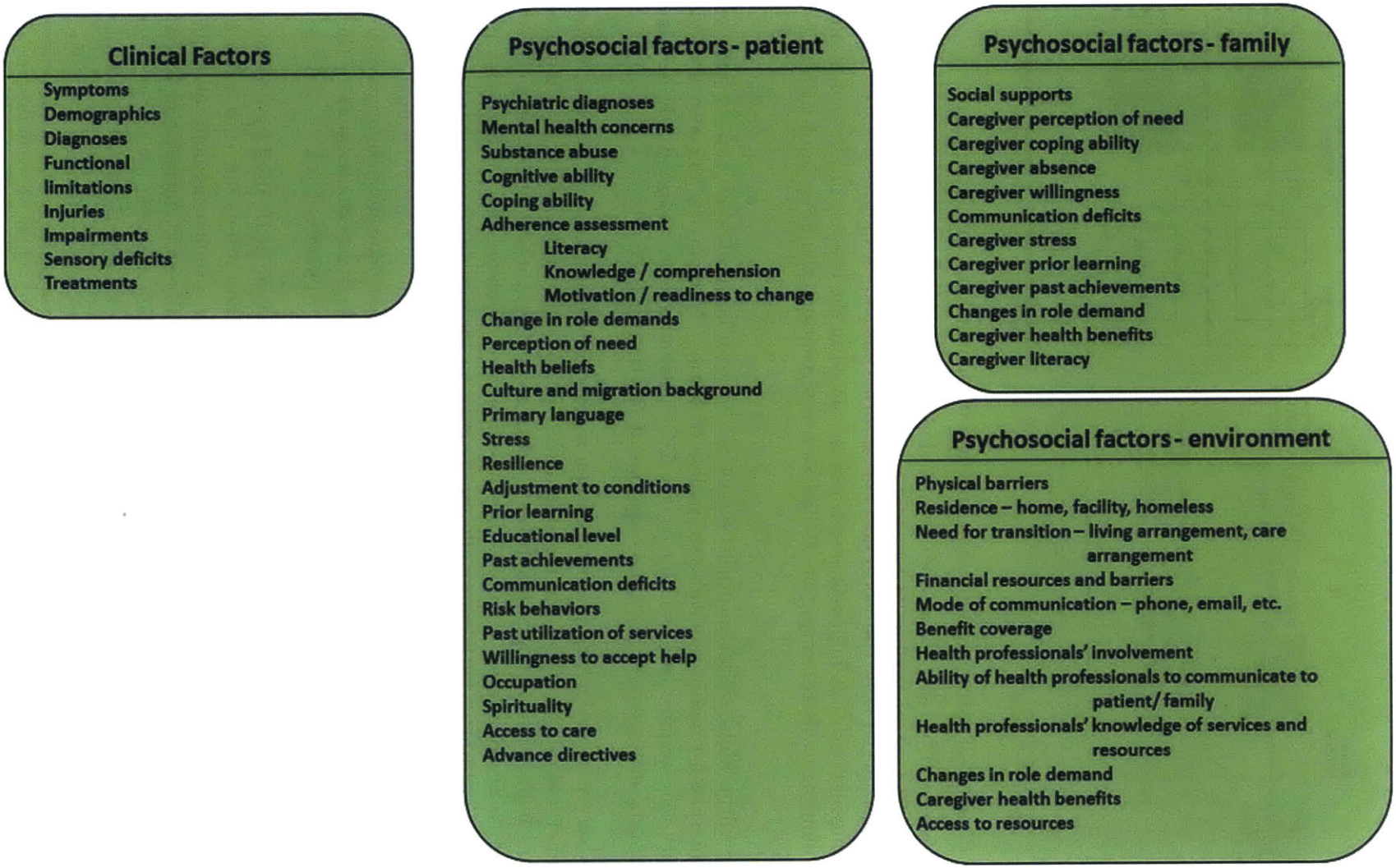


Figure 2-7: Comprehensive consideration of patient needs from a discharge planning perspective - *adapted from [14]*

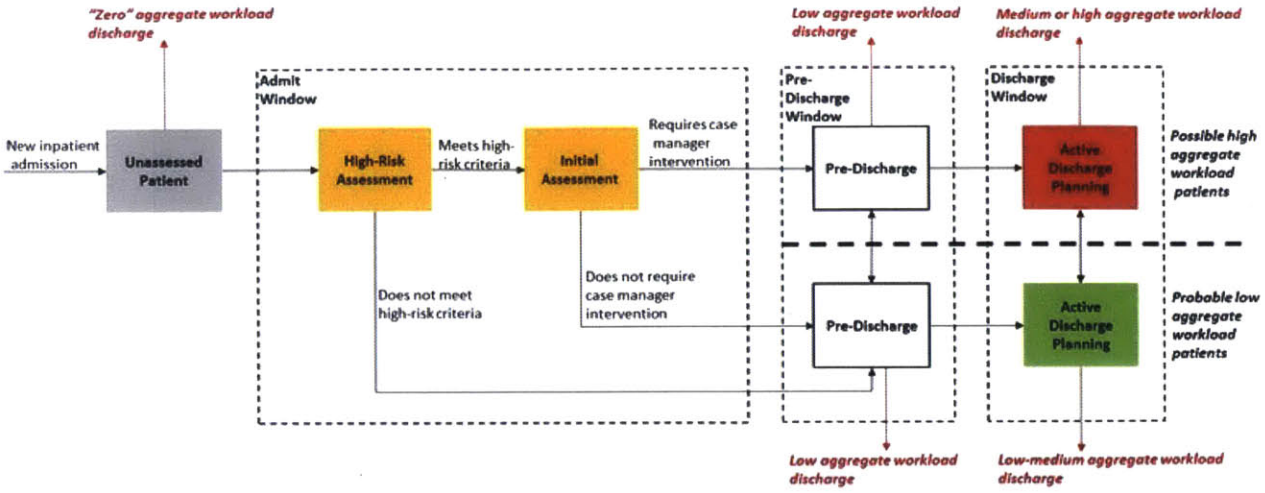


Figure 2-8: Patient progression and flow in the discharge planning plane

either on the weekend or early on a Monday morning. MGH DCP case managers are required to perform a high-risk[3] screen within 24 business hours but this screen is not required for patients admitted and discharged over the weekend. For all other patients a high-risk screen is required within 24 business hours. The high-risk screen can be placed in the context of the typical phases of a discharge planning cycle consisting of: assessment, intervention, planning, implementation, and evaluation[15].

At MGH, the initial high-risk screen varies somewhat between hospital floors and services, but the purpose of the screen remains the same regardless of setting. Namely, the purpose is to identify which patients may require intervention by the DCP case manager. Appendix D provides the high-risk screening criteria used in various settings at MGH.

The high-risk screen is a relatively blunt tool when it comes to identifying the patients who require the highest intervention intensity from case managers. From the DCP perspective, when a patient is identified as high-risk this triggers another type of screening, the initial assessment(IA). The IA is required to be completed within 48 business hours of admission for all high-risk patients. This provides an assessment with marginally more power for identifying which patients will require case management intervention; i.e. the initial assessment can help to identify which patients will contribute to a case manager's workload. The assessment may require examining medical records, meeting with the patient or family, communication with other case managers, communication with other actors involved in the patient's care prior to hospitalization, or any combination of the above. A general template for the initial assessment is also provided in Appendix D.

The template in Appendix D, the HRIA2 template, is only one type of IA used by case managers at MGH. This template contains 88 questions, some free-text and others with drop-down menu choices, designed to answer the question of "Will a patient require case management intervention for discharge planning purposes?" The IA can also take the form of a free-text narrative or follow one of several other templates. Considering the HRIA2 in particular, when it is used not all questions have responses for all patients. The IA may also include a prediction for a patient's ultimate discharge

[3]As with the discussion of acuity in Chapter 1, high-risk is a multi-faceted concept. Because of this, the concept, which may have a very precise meaning in a given context becomes increasingly amorphous as more literature is reviewed[69][11][21][101]. As with acuity we primarily consider high-risk in terms of the workload that cases/patients so designated may impose on case managers.

disposition (e.g., SNF, LTAC, IRF (inpatient rehab facility), home services, etc.).

In addition to providing visibility on the amount of DCP intervention a patient may require and helping to identify any potential barriers to discharge (such as insurance issues), the information from the IA allows the DCP case manager to be a more effective advocate and facilitator for the patient in the intervention phase of case management. To explain, some DCP case manager's use the initial assessment to identify consults and services that the patient may require, such as nutrition, physical therapy, social work, occupational therapy, or speech language pathology, for example. The initial assessment also allows the case manager to be a more active participant and advocate in patient discussions with other members of the care team, such as in multidisciplinary rounds. Finally, the initial assessment can facilitate rapport-building with the patient and/or family that may pay great dividends in later phases of DCP. The degree to which the initial assessment is used for each purpose varies both by patient and by CM / hospital floor.

Together, the high-risk screen and initial assessment comprise the "Admit Window" work in Figure 2-8. Current practice at MGH means that the high-risk screen can be performed relatively rapidly. In fact, the high-risk screens observed rarely took more than 10 minutes. Reportedly, high-risk screens in the past could routinely consume 15-20 minutes and required answering a suite of questions with a template comparable to the HRIA2. By contrast, initial assessments can easily require 25 minutes. Some case managers combine completion of the high-risk screen and initial assessment, while others complete the high-risk screen one day and the initial assessment on a subsequent day. Even with the initial assessment the admit window screens are still a blunt tool, from a positive predictive value perspective, for determining how much work a patient will ultimately require from a DCP case manager. Following the "Admit Window" there are two general categories of patients:

1. Probable low workload patients
   (a) Patients not meeting high-risk criteria
   (b) Patients meeting high-risk criteria but, upon initial assessment, do not require case management intervention
2. Possible high workload patients that meet high-risk criteria and who, upon initial assessment, may require case management intervention

The distinction between probable and possible is used to indicate that, while the positive predictive value of admit window work is low, the negative predictive power[4] is high and can potentially be exploited for modeling purposes when projecting the daily workload for a case manager. To explain, there is a very high probability that a patient identified as not requiring case manager intervention will require a relatively low amount of work from a DCP perspective. Of note in Figure 2-8, information revealed while the patient is in either the pre-discharge state or the discharge window could result in an initially assessed possible high workload patient reclassified as a probable low workload patient, or vice versa. This is not an explicit reclassification but it is taken into account as a DCP case manager prioritizes her efforts.

Referring again to Figure 2-4, sometimes discharge window work coincides with admit window work. This is the case when a patient is near discharge day at the time of the initial assessment. However,

[4]Chapter 6 provides a complete analysis of the negative predictive value of the HRIA in terms or workload imposed over the course of a patients LOS.

for many patients it is possible to identify a pre-discharge period. This is a latent period when the patient's discharge needs are unclear. This is not to say that the patient requires zero work from the DCP case manager, but it is relatively low level work, such as discussions in rounds or, if the patient's LOS extends past a week, weekly update notes on the patient's status and progression relative to discharge. It is possible that the amount of pre-discharge work is higher for patients on floors with a longer average LOS, such as intensive care units or the respiratory acute care unit, but this has not been examined. Regardless, on a daily basis, cases in the discharge window consume most of a DCP case manager's time. Some Mondays are a notable exception to this generally true observation and the reasons and impact of this are discussed in subsequent chapters.

On the one hand, activities for cases in the discharge window can require more time. On the other hand, the recent segregation of CM activities resulting in DCP case managers and UR case managers means that DCP case managers must focus almost exclusively on discharge window patients to support timely and safe discharges. Almost without fail and independent of the specific floor observed, DCP case managers reported that they had to assume more of a reactive rather than proactive posture since the reorganization. Being absolved from UR responsibilities, it is true that DCP case managers have to complete fewer tasks per patient. However, DCP case managers are now responsible for more patients and, at least in the estimation of DCP case managers, the overall aggregate effect is often an increased workload that precludes a proactive posture. Figure 2-9 shows a more granular view of a DCP process map for several common discharge pathways.

The important thing to remember about the process map shown in Figure 2-9 is the overall context in which the processes occur. This context is highly case-dependent and is illustrated by the border encompassing the process map. Four of these factors were discussed earlier in this chapter, but it is clear that floor processes and role boundaries influence both the process map and the workload that a DCP case manager has to complete for a given case.

Depending on the context for the process map, and how this context evolves through time, the DCP case manager may be forced to cycle through processes multiple times. When considering the aggregate DCP workload that has to be completed for a patient, the repetitive cycling can drastically increase the work required for one patient vis-à-vis another patient. What is more, a single factor, frequently but not necessarily insurance, can be the distinguishing factor between a case that requires a disproportionate amount of a DCP case manager's time and one that progresses smoothly until discharge. In addition to repetitive process map cycling there are a handful of patients, less than three percent of the 1600+ cases examined in detail, that would require a highly idiosyncratic process map. Both of these types of patients, those requiring repetitive cycling and those requiring atypical interactions and processes, can result in a high volume of DCP case manager workload during a patient's stay.

Again, similar to each of the decision points in Figure 2-6, each of the decision points in Figure 2-9 implies work for the DCP case manager. We can briefly consider each of the pathways depicted, in turn, to understand some of the key interactions and elements of workload for CM case managers when working on cases in the discharge window. Irrespective of path, DCP case mangers begin active discharge planning when, based on information from the care team (physicians, attending nurses, physical therapists, etc.), the patient's probable discharge needs become clear. This clarity allows the CM to identify options available for the patient upon discharge. The CM then meets with the patient to review options and receive the patient's preferences. It is important to note that patient preferences can, in some instances, be family, health care proxy, or legal guardian preferences

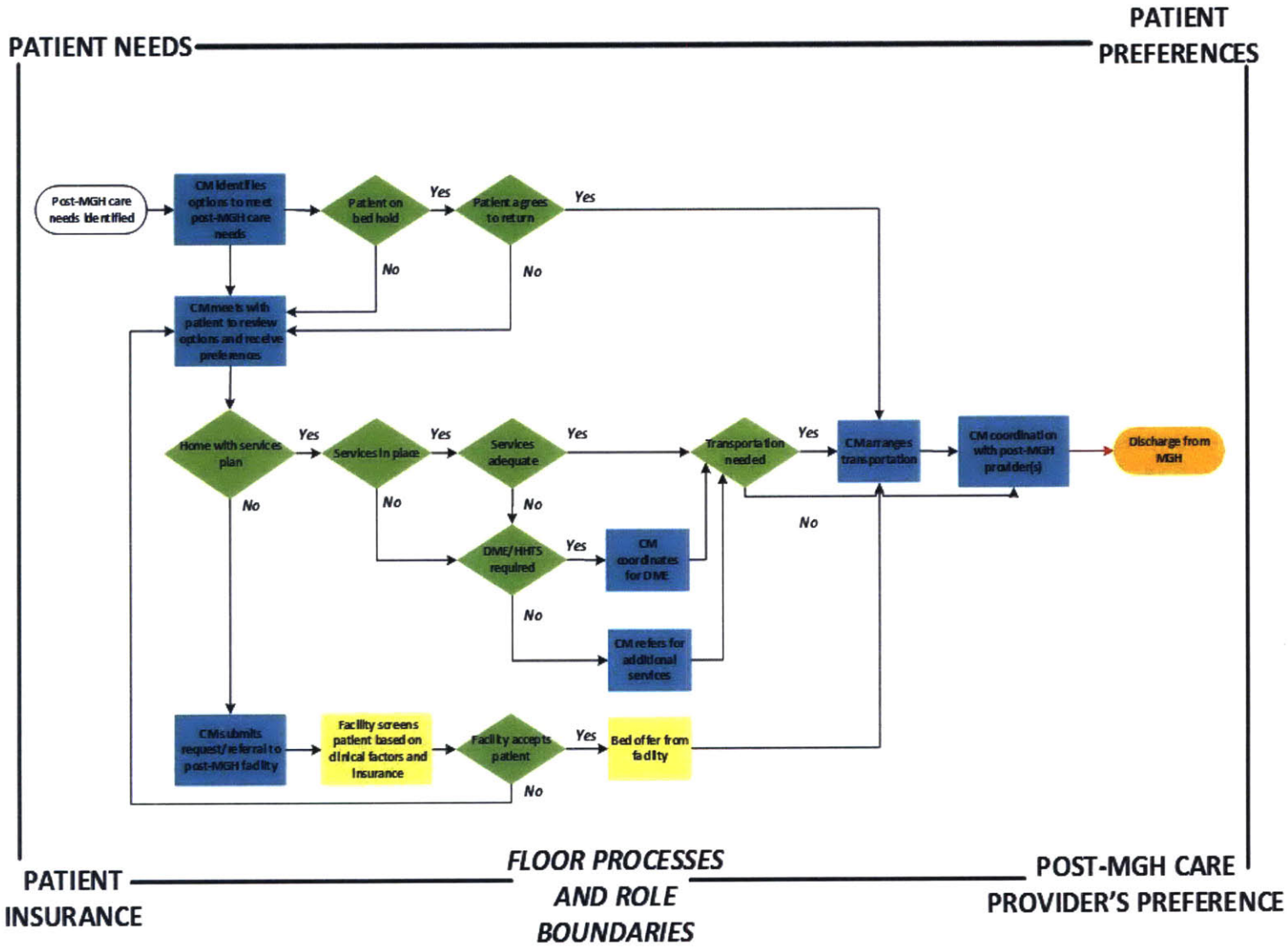


Figure 2-9: Summary process map for several discharge planning pathways

if the patient lacks capacity to make decisions, or the patient has effectively ceded decision making authority to others.

In the topmost path we have a patient that is admitted from another facility. This patient is a long-term care resident at another facility and may have a statute-mandated bed hold or an effective bed hold. For example, a patient with a MassHealth policy admitted to MGH from a SNF has a 20-day bed hold and may return to the admitting facility when ready without requiring additional facility screening or insurance authorization. A private-pay patient may also have a bed hold. For a patient without a statute-mandated bed hold, a patient may have an effective bed hold where the admitting facility has agreed to accept the patient back when ready for discharge from MGH. If a patient on a bed hold agrees to return to the admitting facility then the discharge planning process is relatively straightforward. The case manager may merely need to coordinate transportation for the patient to return to the admitting facility. However, the patient may not wish to return to the admitting facility and this could require additional work for the CM.

In the next path, where a home plan is appropriate, a patient may already have care services in place. If these services are appropriate for the patient upon discharge (such as visiting nurse association or VNA services) then restarting the services requires coordination with the home care provider agency and may be relatively simple. Alternatively, a patient may require additional home services or durable medical equipment (DME, such as home oxygen, nebulizers, tele-monitoring, hospital bed, hoyer lift, etc.) and this requires additional coordination on the part of the DCP case manager with post-MGH service providers. Depending on the patient's insurance and needs the amount of work this coordination requires can be extensive. What is more, a discharge to home with multiple services can place an intense burden on a DCP case manager on the day of discharge. For example, a patient requiring home infusion services (e.g., IV antibiotics) upon discharge may require coordination with family and multiple providers, in addition to coordination with the care team to modify a patient's infusion schedule, all within a limited timeframe[5].

The bottommost pathway in Figure 2-9 shows a summary pathway for patients requiring a post-acute care facility for which there is no bed hold. Again, the process begins with a patient meeting to receive preferences for a referral to a facility. These preferences are then typically relayed to CMRS personnel who enter the referral into an IT system supporting the centralized referral process at MGH, 4Next. The referral contains clinical information for the patient. The post-acute care facility then performs a clinical screen of the patient to make sure they are at the appropriate LOC for the facility. If a patient passes the clinical screen the facility seeks insurance authorization for the patient. Insurance authorization could fail if, for example, the patient's insurance is not contracted with a post-MGH service provider. If authorization is obtained then the facility accepts the patient and offers a bed. Acceptance and a bed offer may be concurrent or separated by a span. When a patient is accepted to a facility and has a bed offer the case manager helps coordinate transportation to the facility upon discharge. Of note, certain post-MGH facilities are highly desirable to patients. These facilities always have a demand which exceeds their capacity and they can be very selective in accepting patients. In this context patients at MGH can impose an additional burden on case managers by demanding to be referred to one of these high demand facilities and only this facility, even if the patient is not an appropriate candidate for placement in the high demand facility. A resulting, often inevitable, denial can also cause the DCP planning process in Figure 2-9 to restart.

[5]The DCP case manager frequently interacts with MGH's Infusion Resource Center (IRC) and New England Life Care liaisons to coordinate for home infusion therapies.

Obviously, the process map shown in Figure 2-9 is far from exhaustive and fails to capture the potential complexity in a DCP case manager's work. However, the process map does illustrate some of the key elements of a DCP case manager's discharge window workload such as:

- Meeting with patients, families, members of the care team, and post-MGH provider liaisons
- Phone calls to families, post-MGH provider representatives, insurance case managers, and potentially many more entities
- Faxing required documentation
- Coordinating transportation for a patient

The DCP case manager also has documentation requirements, such as Face-to-Face forms(F2F) for certain VNA patients or, for example, extensive statute-mandated long term care forms for MassHealth patients requiring LTAC placement upon discharge from MGH. Additionally, the case manager may have to interface with MGH's Patient Financial Services (PFS) to overcome insurance-related barriers to discharge. A case manager's work can be very complex and distilling a case manager's workload into discrete elements such as phone calls and meetings is not meant to understate this complexity. Instead, identifying elements of a DCP case manager's workload is a key step in quantifying workload, both on a patient-by-patient basis and in the aggregate. In summary, the DCP case manager has to potentially interface with many people to develop and execute a safe discharge plan and the quantity, content, and sequence of these interactions is highly dependent on patient needs, preferences, and insurance.

2.4 MGH case management organization and resource allocation

The inpatient CM organization at MGH consists of a leadership team, two clinical nurse specialists, DCP case managers, UR case managers, a 12-person CMRS support staff, and other support personnel such as a quality improvement individual and a special projects officer. During the period of this work the organizational FTE was 70-72 apportioned among approximately 100 people.

The leadership team consists of an executive director for case management and two nursing directors for case management. Each nursing director manages approximately 50 people and is responsible for case management on a number of floors and associated services at MGH. For example, one nursing director manages psychiatric, medicine, and oncology floor case managers (among others) while another manages surgery and neurology floors (also among others). Each of the nursing directors also oversees other areas of MGH case management or case management initiatives, such as payor-side activities or the centralized referral process.

The clinical nurse specialists play a role in CM education and improvement initiatives and are also highly active in very difficult cases. These individuals are also available to perform UR or DCP activities as required in covering for case managers and in crisis situations such as "Code Disaster" when the capacity of MGH's ED is strained and the focus of everyone becomes discharging inpatients so patients in the ED can be moved to other floors and thereby free up space in the ED.

Some of the roles of the CMRS personnel have been recounted earlier. Considering UR, these personnel help to mediate communication between UR case managers and insurance companies. On

the discharge planning side CMRS personnel also perform a variety of activities to support case management at MGH. These activities include entering referrals for patients in 4Next, delivering required Medicare messages to patients, collecting and ensuring completeness of documentation such as F2F forms, among others. Seven of the CMRS personnel are distributed among 45 floors to help support both DCP and UR activities. The extent of these activities varies based on the preferences of respective floor case managers and the skills of the assigned CMRS individual. Some CMRS personnel assist in performing patient assessments and arranging transportation. During our work there was an effort to standardize the role of CMRS personnel taking place in the context of a larger initiative to more precisely identify, define, and standardize roles and role boundaries among care team members.

The face of case management at MGH is the front-line UR and DCP case managers. Appendix E shows how the case managers were distributed across 45 floors at MGH to perform DCP and UR on a weekday. Appendix E also indicates the number of beds that a case manager is responsible for. During our research there were 39 DCP positions, 17 UR positions, and 2 combined positions for 58 positions total. On a typical workday, depending on case manager absence because of vacation, sickness, or other reasons, 52-55 case managers would be filling these 58 positions[8][6].

Most of the DCP and UR positions are permanent positions, meaning that they had a permanently assigned case manager who, if available for the day, would fill that position. Three of the DCP positions were manned by members of the "float" pool of case managers. These "float" pool case managers also provide coverage in the absence of permanent case managers and provide weekend coverage. The weekend coverage typically consists of five or six (and more rarely seven) DCP case managers. On the weekend the focus is on "last mile" facilitation of weekend discharges that have, usually, been in large part coordinated by weekday case managers. One CMRS individual is also active on the weekend. 15 of the UR positions are permanent positions. With the implementation of MGH's dyad model there are in fact UR teams that are responsible for a pool of floors and cases.

Table 2.1 shows some case:case manager ratios for select floors at MGH. The ratios vary by type of floor and between DCP and UR case managers. The overall benchmark caseloads that MGH is aiming for are approximately 25:1 for DCP and 50:1 for UR. On the DCP side some ratios are lower, particularly for the complex patient population on a Neuro floor like Lunder 7 (general Neuro floor, not the Neuro ICU) where the DCP ratio is 19:1.

As indicated in Appendix E (positions labeled A,B,C, and D) there are other positions filled by case managers. The psychiatric consult case manager works to facilitate discharge of patients, requiring an inpatient psychiatric admission, from an MGH floor to either Blake 11 at MGH or another facility providing acute psychiatric services. The late day bed offer case manager helps to facilitate discharge of patients when a bed offer is received, as the name would imply, late in the business day. The Ortho/RAPT case manager works on an on-call basis to evaluate elective orthopedic patients using MGH's Risk Assessment and Predictive Tool (RAPT)[76] to predict discharge needs pre-surgery and, based on the RAPT score, pre-coordinate discharge needs (e.g., post-acute facility, VNA, etc.). Finally, the SNF waiver CM helps facilitate discharge of select Medicare patients who have not met the "three consecutive midnight rule" during their inpatient stay but are ready for discharge to a lower level of care.

Table 2.1: UR and DCP caseloads (bed responsibility) for select floors at MGH

Building	Floor	Service	Beds Assigned and Position Number	
			Discharge Planner	Utilization Review
Blake	11	Psychiatry	24 - Position 8	24 - Position 8
Gray/Bigelow	9	Respiratory acute care unit (10) / Medicine (8)	18 - Position 13	65 - Position 5
Lunder	6/7	Neuro ICU / Neuro	28 (22/6) - Position 29	65 - Position 5 / 64 - Position 16
Lunder	7	Neuro	18 - Position 30	64 - Position 16
Lunder	7/8	Neuro	19 (8/11) - Position 31	64 - Position 16
Lunder	8	Neuro	21 - Position 32	64 - Position 16
Lunder	9	Oncology	*22 - Position 33*	64 - Position 17
Lunder	9/10	Oncology	22 - Position 34	64 - Position 17
Lunder	10	Oncology	20 - Position 35	64 - Position 17
White	8	Medicine	24 - Position 38	82 - Position 2
White	9	Medicine	*25 - Position 40*	65 - Position 5
White	8/10	Medicine	22 (2/20) - Position 39	82 - Position 2
White	11	Medicine	24 - Position 41	49 - Position 8

Example Key: Discharge planner case manager position 29 is responsible for 28 patients total, 22 on Lunder 6 (Neuro ICU) and 6 on the Lunder 7 Neuro floor. These 28 beds are followed for utilization review purposes by utilization review case managers in position 5 (22 beds on Lunder 6 Neuro ICU) and position 16 (6 beds on Lunder 7 Neuro floor). Utilization review case manager position 5 is responsible for 65 beds total and utilization review case manager position 16 is responsible for 64 beds total.

2.5 Some sources of case manager workload variability

Shifting the focus more explicitly to discharge planning, it is possible to identify a number of potential sources of workload variability, both for a given case manager on a day-to-day basis and between case manager positions. A similar assessment could be completed for UR case managers or between UR and DCP case managers. In fact, subsequent future work practically necessitates an assessment between UR and DCP case managers to determine if the current benchmark caseloads are appropriate. What follows is an essentially qualitative assessment to provide context for the problem statement presented in Section 2.7. This qualitative discussion is supplemented by a quantitative assessment in Chapters 4-6.

One potential source of workload variability for a given case manager centers on daily census variability; i.e. how many of the beds for which a case manager is assigned responsibility are occupied on a given day? However, the acute care capacity of MGH experiences a relatively consistent and high level of utilization meaning that, ultimately, this potential source of workload variability pales in comparison to other sources.

Another source of workload variability centers on the differential weekday and weekend staffing patterns for case managers at MGH. In fact, this variation is acknowledged by all in the CM organization at MGH, so much so that our observations were confined to other weekdays so as not not disrupt case managers on potentially extreme high workload days. On the discharge side, much of the work for weekend discharges is completed on Friday so that the minimal DCP staff on weekends can facilitate weekend discharges. Furthermore, even if a patient does not appear to be nearing discharge on Friday afternoon, or if discharge needs are still unclear, in the intervening 60 hours the patient could progress so that the patient is ready for discharge on Monday. This type of workload variation can be particularly insidious because, even if the case is not particularly complex, work that could be completed over a multi-day span without delaying a patients discharge may have to be compressed into fewer hours. Considering admissions, because the reduced weekend staff typically focuses on discharge window activities, the weekday case manager could be faced

with a large number of patients requiring high-risk screening and initial assessment on Monday.

Yet another source of workload variability for a case manager can be traced to variations in patient needs over some timeframe even given a relatively constant census. As explained, some patients have much more extensive discharge planning needs than others (inter-case variability). This explains workload variation between patients but, in the aggregate, a case manager could have an unusually high number of patients with extensive needs or, conversely, an unusually high number of patients with virtually no needs. Similarly, the average patient on some floors may have less needs than on other floors (inter-floor variability). For example, a patient undergoing an elective orthopedic procedure may have a more well-characterized discharge path. The discharge needs may be known with a large degree of certainty even before admission and because elective procedures are accompanied by insurance pre-authorizations, insurance issues are unlikely to be a major driver of workload. The differing benchmark caseloads between floors and discharge planners represents an explicit acknowledgement that some patient/floor populations typically require more of a discharge planner, over long time horizons.

There is a fundamental source of daily workload variability that encompasses all of the aforementioned sources of variability; namely, "how many cases in a case manager's assigned census will require case manager intervention today[6]"? That is, a case that is not active requires, in essence, "zero" work for the day. The key to predicting the daily workload for a case manager hinges on predicting the "active census" for that day. Of course, predicting whether any given case will be active today is difficult, but, given a pool of patients, each of which may be at a different point in progression in the discharge planning plane (as illustrated in Figure 2-8), it becomes possible to predict a case manager's workload for the day, replacing the prediction of whether a particular case will be active today with a prediction of aggregate workload given the number of unassessed, pre-discharge, and discharge window cases[7]. This is the modeling approach taken in Chapter 7.

2.6 Data sources

The primary data sources used in our research and analysis are briefly outlined below. The relative importance of available data sources changed over time as the focus of the project shifted from predicting high aggregate workload patients to predicting a case manager's daily workload. Appendix A takes the form of a concept paper that outlines additional data sources and how they can be used to develop a model to predict high aggregate workload patients. In essence this tool would be a more refined version of the current high-risk screen and initial assessments. This appendix also outlines some of the problems with the existing data and the caution that must be exercised when developing a tool using this data. Of note, some of the data sources described below, particularly those derived from the Morrisey case management system, may not be available for future patients given the MGH-wide implementation of Epic.

Morrisey-derived data sources

Case Manager Notes - This is **the** primary data set used for our work. As a case manager performs

[6]The concept of an active census is introduced formally in Chapter 5 and discussed extensively, in terms of predicting daily case manager workload, in Chapter 6.

[7]See Chapter 5 for a formal definition of these terms.

DCP activities for a patient this work is recorded in the form of case manager notes. As explained in Chapter 4, this allows manual counting of workload events and it also provides the text for which automated and semi-automated text analytical techniques are developed. The note is originally written within the Morrisey platform and then copied to another system, eBridge, where it is viewable by other members of the care team during a patient's stay. These notes are then uploaded to CAS. It is from CAS that the notes were retrieved on behalf of MGH case management by the MGH Lab of Computer Science. The notes have the associated text, a timestamp, note author and a header that has limited discriminatory value, being either "CM Initial Evaluation" or "CM Progress Note"[8].

Dyad CM Workload Report Raw Data - The raw data for this report contains the number of discharges from a given floor, whether a patient was screened for high-risk criteria and if a patient met these criteria, and whether any answers were recorded for an HRIA2 initial assessment.

The rather circuitous method by which the notes are uploaded to CAS can be problematic. The timestamp of the note indicates when it was copied into eBridge. This means that, in general, only the date of the note, and not the time is reliable. There is also a delay between when case manager work occurs and when it is documented, though this is not a major drawback. More problematic is the chance that some notes are missing from the CAS data pull because a Morrisey note was not copied to eBridge. Of course, in some cases it is possible that "missing" work may not have been documented. It is easy to infer the type[9] of missing work by combining the information from other data sources, particularly the 4Next–derived centralized referrals, the patient's discharge disposition, and, for admit window work, the high-risk screen and IA summary data (i.e., whether a high-risk screen or IA was completed for a patient but not the accompanying text) in the Dyad CM Workload Report raw data. Preceding and succeeding notes for a patient are also indicative of missing notes and their general content. This aspect of missing work/notes is revisited in Chapter 4 where it will be shown that the problem is not insurmountable.

The case manager notes also contain useful information on patient psychosocial factors affecting patient discharge and DCP workload that may be hard to obtain from other sources. This information is frequently found in the text of initial assessments or in notes detailing why a patient was denied by a post-MGH facility. Finally, the case manager notes are key to identifying when a patient transitions from one phase on the discharge planning plane to another phase.

Morrisey Level of Care Raw Data – This data contains general discharge disposition information for a patient (e.g., IRF, SNF, home with VNA services, etc.), as well as specific identification of the post-discharge care providers for a patient. It also allows a determination of the number of post-discharge providers for a patient.

4Next-derived data sources

CMSU Referrals Raw Data - This data source records when referrals for a patient are made to certain facilities such as SNFs, LTAC facilities, and IRFs. As alluded to above this data can be useful in identifying potential missing notes and in helping to identify when a patient enters the discharge window.

[8]A major part of the text-analytical techniques described in Chapter 4 involved further discrimination of note types based on textual markers.

[9]Inferring the amount of missing work is subject to more uncertainty.

Epic-derived data sources

General EPIC Data for Hospital Encounters - This source provides a wealth of data on patients, including discharge disposition, admission source (e.g., admission from a facility, self-referral from home, etc.), hospital inpatient admission and discharge times, ED departure time (if applicable) and first inpatient department.

Admission, Transfer, and Discharge Data - This data contains timestamps allowing determination of when a patient arrived to and departed from an inpatient floor. The data also provides the room where a patient transfers to or from. From the timestamp and room information it is a straightforward matter to determine the census on any given day for a case manager.

EPSi-derived data sources

General EPSi Data for Hospital Encounters - Similar to Epic, EPSi provides a number of data points for a patient. EPSi contains information relevant to patient billing such as payor, secondary payor and contract family for a patient. EPSi also provides the admission and principal ICD-9 diagnoses for a patient and the diagnostic resource group (DRG) codes for the patient. The principal procedure code for a patient is also available. In addition, EPSi provides the hours since a patient's last outpatient and inpatient visit. This type of data has to be considered carefully because it does not necessarily provide accurate information for a patient that may have been seen at another facility. This data does allow determination of how many inpatient and outpatient visits a patient has had at MGH over a given time frame, or within a period of time before and/or after an admission.

EPSi ICD-9 Diagnosis Codes - A patient can have up to 12 different ICD-9 diagnosis codes for a single claim so many patients have more than just an admission and principal diagnosis. This data source contains additional diagnosis codes for a patient.

Of course, interviews with case managers and case management leadership are also key data sources; these interviews. along with observation, allowed us to map the problem space and develop our problem statement.

2.7 Problem statement and overarching approach

MGH CM leadership was very interested in working with the MGH-MIT Collaborative to address some of the issues they perceived with their current staffing scheme. CM leadership had received feedback from case managers, particularly DCP case managers, questioning the distribution of workload between UR and DCP. Some of these concerns predated the transition to MGH's dyad model of case management but, as described above, DCP case managers indicated in interviews an increase in perceived workload since implementation of the dyad model. These case managers also expressed a belief that cases were becoming more complex in terms of psychosocial factors that had to be considered for discharge planning. Case management cast the problem in terms of a resource allocation problem and possibly working to develop an analog to acuity systems used by staff nurses. This acuity tool would then be used to develop a staffing grid that would result in a more equitable distribution of workload.

Working as part of the MIT-MGH Collaboration, and with CM leadership support, we ultimately

formulated the problem from a slightly different perspective. Specifically, the problem statement can be summarized as:

> The current static staffing scheme, based on the number of beds a case manager is responsible for (caseload), cannot effectively address the variability in daily workload encountered by DCP case managers.

Given the above problem statement, the primary goals of the project crystallized into:

1. Developing a methodology to measure and predict CM workload in an effort to inform staffing decisions
2. Developing candidate staffing schemes that incorporate the flexibility required to effectively address variability in a case manager's daily workload and/or reduce observed variability

In the context of the problem statement and project goals, the overall approach to the project centered on overcoming the main challenges to goal realization. The fundamental challenge stems from the fact that it is not immediately clear how to quantify workload in a meaningfully measurable manner; i.e., in a way that reflects the workload experienced by case managers. The secondary challenge is, assuming that the fundamental challenge of meaningful measurement can be solved, identifying features, either at the patient-level or at an aggregate level, allowing reliable prospective prediction of a case manager's daily workload. Chapter 4 details the approach taken to develop a meaningful workload metric and Chapter 5 provides a description of the current state of DCP workload distribution and variability, both at the patient-level and daily level using this metric and derived metrics. Chapter 6 presents a predictive workload model based on the metric and current state analysis that can be used to help inform staffing decisions. In Chapter 7 we present preliminary operational recommendations, as well as recommendations for future work.

Chapter 3

Literature Review

This chapter provides a targeted literature review covering three primary areas:

1. Systems for measuring case manager workload (Section 3.1)
2. Text analytical techniques applied to unstructured (free-text) healthcare data with varying response variables (Section 3.2)
3. Machine learning techniques with applications for imbalanced data sets (Section 3.3)

There are efforts at developing systems to measure case manager workload described in the literature. However, even though eight years have passed since the Case Management Society of America's (CSMA) caseload concept paper, two observations from that paper remain salient from the perspective of a literature review:

1. The literature on measuring case manager workload and determining appropriate benchmarks for caseloads is relatively sparse
2. The quantitatively rigorous research findings available in the literature are confined to either narrowly defined clinical areas or specific "in-house" programs[14]

As a result of the second of these observations, the utility of the existing literature applied to the MGH inpatient case manager setting may be limited; as explained in Chapter 1 and Chapter 2, the term "case management" can refer to different roles, processes, and work elements depending on the setting.

Still, the elements necessary to develop an effective methodology for measuring and predicting case manager workload are identifiable in the literature. Often these elements are only qualitatively discussed, as statements of the form "when determining appropriate caseloads factor x, y, and z should be considered" are common, but these types of statements still informed our work.

The biggest deficit in the literature, at least from the perspective of our work, is an under-consideration of the fact that determining appropriate benchmark or baseline caseloads for a case

manager position is a fundamentally different problem than making an operationally useful prediction about the amount of work required of a case manager on a given day. Time horizons or, more correctly, the timing of work matters.

The existing literature focuses on providing an answer to the question of, for example, how many cases should a case manager on a general medicine floor be responsible for as compared to a case manager on an oncology floor? The mechanism for answering this question usually takes the form of assigning an acuity score to cases and comparing the acuity mix of cases, over an extended time horizon, between floors. However, an acuity or work score used in this manner, while it may be closely correlated with the amount of work that must be completed for a case over a patient's entire LOS, says little about how much work must be completed for a case, or by a case manager, on any given day.

The concept of a benchmark caseload is incomplete because of its reliance on average or "typical" workloads. Case managers experience the actual workload on any given day, a workload that exhibits marked variability. It should be obvious that "appropriate" benchmark caseloads can still result in a vastly suboptimal allocation of case management resources, the most important of which is frontline case managers, on any day that all case managers do not experience their average daily workload. The most important contribution of our work to the body of literature reviewed below is in more fully considering the timing of work for a case, thereby allowing a dynamic element for resource allocation to be introduced.

Our literature review differs slightly from literature reviews as commonly constituted. Rather than simply making the reader aware of existing literature related to our work, we try to point out the relationship of the literature to our work, as well as what we view as "gaps" in the literature from the perspective of our work. We also endeavor to identify linkages between the literature and our work, including how existing approaches or techniques were extended or applied in this thesis.

3.1 Measuring case manager workload

As Cesta asserts, suboptimal staffing ratios are usually the primary reason case management departments do not function as well as they could[13]. As explained in the introduction to this chapter, staffing ratios can be understood to mean baseline caseloads benchmarked to account for the factors that drive case manager workload for a case and the differential prevalence/impact of these factors in different patient populations (different hospital floors). Cesta further identifies eight overarching factors influencing case manager workload:

1. Role function
2. Model design (e.g., dyad, triad, etc.)
3. Payor mix
4. Intensity of service
5. Complexity of patients
6. LOS of patients
7. Staffing patterns

8. Use of technology[13]

This list of factors, though more general, is essentially identical to the CSMA factors identified in [14], a portion of which are provided in Figure 2.3. Cesta's factors 3, 4, and 5 are also extensively considered in the high-risk screening and initial assessment used by MGH's inpatient case managers (see Appendix D) In fact, this phenomenon is a common feature in case management literature – there is broad agreement on the factors that drive case manager workload. Despite agreement on the factors driving case manager workload, developing a system to meaningfully weight these factors as a prerequisite for assigning a score to a case has proven more elusive.

Huber and Craig present a system, the CM Acuity Tool, for scoring cases based on the concepts of acuity and dosage in a series of three articles[82][81][60]; Craig extends this work in [59]. This system was developed in a large, telephonic case management company setting and begins with a list of factors that drive workload for a case, a list not unlike the lists described above. In a series of steps these factors are mapped to case complexity and this complexity is mapped to an acuity score (1-5) correlated with the amount of work required for a case.

One provocative facet of this work is the method of summing individual case acuities to get a weighted aggregate caseload for a case manager []. The rationale is that this weighted caseload more accurately reflects the work required for a set of cases than a simple count of the number of cases a case manager is responsible for. Also, the system developed by Huber and Craig allows case managers to score cases at different points in the case manager service delivery cycle[82]. Thus, a case initially scored as a 5 could conceivably be scored as, for example, a 3 at some later point in the service delivery cycle. Presumably this would occur because work done by a case manager resolved some of the complexities surrounding a case meaning less complexity and less future work remains for a case.

In general, case management acuity scores relate, however loosely, to the total amount of work remaining for a case and say little about the work required for a case on any given day. This is a key difference between case management acuity systems and, for example, nursing acuity systems that attempt to score patients daily. To explain, if a patient is still hospitalized, then it can reasonably be assumed that that patient has clinical needs that imply required work for an attending nurse[1]. This assumption does not hold from the perspective of DCP case managers at MGH.

The setting in which Huber and Craig's work was completed is different than that of MGH. Still, the concept of a weighted caseload or census can be modified to be more appropriate for the MGH setting. In fact, we use an active census concept in acknowledgement of the fact that in the MGH setting the daily workload for a case manager is driven, in large measure, by the number of cases active on a given day, irrespective of specific case factors related to complexity. We further introduce a weighting component by segmenting cases according to where they exist on the discharge planning plane. This is covered extensively in subsequent chapters but this method effectively deals with some of the issues related to a factor-based case or caseload score. In particular, there is an implicit assumption in factor-based scores that complexity maps neatly to workload; this is not always true. Even when the mapping between complexity and workload holds, the relationship is more between complexity and total workload for a case during a patient's LOS, not the workload required on any given day of the LOS.

[1] Although the works cited in Chapter 1 suggest that different systems of quantifying workload can yield materially different results.

Balstad and Springer describe efforts to quantify case management workloads in a setting more similar to MGH[38]. In many respects, judging by the number of times their work has been cited, this work is not adequately appreciated; in fact, this work seems relatively obscure. In addition to outlining a method for quantifying case management workload, Balstad and Springer correctly identify why quantifying case manager workloads is vitally important.

As alluded to in Chapter 1, the motivation for this project can be considered in the context of a "morale" problem. When case manager workload is demonstrably not distributed equitably it is natural for those who feel that they are shouldering more than their "fair share" of the burden to question staffing policies. Now, those with a moralist bent, who have likely never experienced the extreme levels of workload possible for some inpatient case managers at MGH, will inevitably counter that this is an "unprofessional" perspective to take. Yet, this line of reasoning completely misses the mark. First, any feeling by case managers that workload in not distributed equitably does not prevent them from faithful and tireless execution of their duties. Furthermore, any complaints stemming from an unequal distribution of workload are rooted in the desire for case managers to perform the best job possible on behalf of patients. Consider, the most common complaint we heard when interviewing and shadowing case managers was the lack of time to spend with patients[7]. The real source of complaint about any workload inequity has little to do with the amount of work required of case managers; the primary source of any complaints concerning unequal distribution of workload, whether expressed or internalized, is that less effort can be devoted to each patient.

Balstad and Springer speculate, just as we speculate, that an inequitable distribution of workload may result in patient outcomes that are less desirable than can be achieved with a more equitable distribution of workload. As they suggest, with a method of quantifying case workloads and, more specifically, the amount of case manager work for a specific case, it should be possible to track patient outcomes as a function of the work performed by a case manager for a patient/case. Now, Balstad and Springer do not explicitly suggest how patient outcomes should be quantified, but several response variables naturally suggest themselves such as, for example, inpatient admissions or cumulative inpatient days, during some time period following discharge.

Several studies have looked at the impact of individually-tailored discharge planning, as representative of the discharge planning performed at MGH, versus more standardized discharge planning on patient outcomes. One of the most comprehensive of these studies is that authored by Goncalves-Bradley et al.[74]. This study is periodically updated and as of 2016 looks at outcomes for 11,964 participants in 30 trials. The authors conclude that the effort expended to develop discharge plans specifically tailored for the individual bring about a small reduction in hospital LOS and three-month readmission rates for older patients. The authors further conclude discharge planning may lead to increased satisfaction, both for patients and healthcare professionals, while there is no significant evidence individually-tailored discharge planning reduces overall costs in the health system.

Looking more deeply at the results of the study suggests that the impact of individually-tailored discharge planning may be understated. Though a "small" reduction in LOS is reported, this reduction has a magnitude of 0.79 days. We feel that this is a noteworthy result. Even granting that the study found minimal impact on individually-tailored discharge planning for reducing overall costs in the healthcare system, this reduction in LOS represents a significant increase in accessibility to healthcare services. Admittedly, tracing patient outcomes to a specific intervention, such as individually-tailored discharge planning, is a task fraught with potential pitfalls. However, if the amount of work completed for each case by a discharge planning case manager can be quantified,

tracing the impact of case manager work to patient outcomes becomes a more tenable proposition. As implied in Balstad and Springer's work, this can form the basis for a powerful business case to be made by a case management department when making the argument for increased funding, or at least a static level of funding, in an increasingly fiscally constrained environment. Ultimately, making the business case for resources is one of the primary functions of an acuity or workload quantification system. The basic idea supporting an equitable distribution of case manager workload is that equity always allows a similar amount of time to be spent on cases where similar needs are present.

Balstad and Springer's work also hints at a deficit in many types of recent healthcare scholarship, with the exception of scholarship emerging from the econometric realm. The increased focus on the patient as the focal point of any healthcare delivery system has proved to be a valuable paradigm shift when it comes to evaluating any proposed changes to this system. However, this focus on the patient is frequently almost to the exclusion of other actors in the system. That is, in terms of the long-term performance of a healthcare delivery system, healthcare professional satisfaction and outcomes have to be placed on essentially equal footing with patient satisfaction and outcomes. Healthcare professional burnout in general, and case manager burnout in particular, are significant factors to consider when developing any type of staffing scheme[32]. Proper workload quantification helps to guard against conditions that increase burnout among case managers.

As we investigated the available literature, we faced a dilemma similar to that faced by Balstad and Springer nearly a decade ago – an exhaustive literature review revealed no instrument to quantify case manager workload or measure patient acuity in a case management setting similar to MGH. There is an established history of developing and using acuity systems to measure nursing workload with the aim of predicting required staffing levels across different patient populations and shifts[89][33][46][47][64][103][66][83][100][105]. A review of the literature suggests that the most successful employment of these systems occurs in an ICU setting[48][98][33][89][46]. This is not surprising given that the daily workload required for a patient requiring an ICU level of care, while subject to variation, is likely subject to a narrower range of daily workloads. The very fact of an ICU-setting becomes an important contextual delimiter that facilitates prediction of the daily nursing workload associated with a case. While none of the systems for quantifying nursing workload were directly applicable to our work, the basic techniques used to develop these systems do have some cross-over applicability. Our basic premise was to identify discrete tasks that may be completed by case managers and then, in consultation with case managers and via direct observation, assign a score, or range of scores to discrete work events. Chapter 4 provides a detailed description of our scoring methodology.

While our system, though relying on case manager input, did not use a formalized Delphi technique, Balstad and Springer's quantification system made extensive use of the well-established Delphi technique. 34 patient need categories, later reduced to 26, were identified and case managers were consulted to develop an acuity score for each patient need category. It bears noting that these patient need categories are very similar to the factors considered by inpatient case managers at MGH during the high-risk screen and initial assessment. The end result was a four-level acuity system in which a score was assigned to each patient need category. Considering the interclass correlation coefficient (0.888) or the Kappa statistic (0.411, p-value $<.0001$) suggests high intra- and inter-rater consistency, respectively, using the Patient Acuity Case Management Evaluation (PACE) tool resulting from the Balstad and Springer's research. This consistency is noteworthy but, again, the scores from the PACE tool are most closely correlated with the total amount of work

completed for a case over the course of a patient's LOS.

The PACE Tool is similar to some of our earlier efforts to elicit rankings from case managers on the amount of work required for a case with certain characteristics. The results of these efforts are instructive and point to the difficulty when considering case factors in any type of additive way to form a case score, either acuity-based or work-based. We consulted with case manager clinical specialists, individuals that are frequently enlisted to facilitate discharge planning for particularly difficult cases. We presented two case manager clinical specialists with a list of 46 case factors / patient need categories and asked them to rank the factors on a five-point scale. Almost every ranking, with the exception of cases with guardianship issues, was provided with a qualification of the form, for example, "this factor usually results in a case that is a 3, but if factor x and y are also present the case would be a 5, while if these other conditions hold the case would actually be a 1."

Again, with a solely feature-based scoring system complex interactions among the factors determine the case manager workload a case presents. These factors are not necessarily additive. As a contrived example, consider a case with a recalcitrant patient, with no insurance coverage, and a lack of available placement options if placement in a sub-acute facility is required. If this patient is safe to discharge to a home setting then the case manager work could be very minimal. On the other hand, if the patient requires placement in a long-term acute care facility then the work required could be extensive. Put another way, sometimes $1 + 1 + 1 = 0$, where 1 indicates the presence of a case factor. Other times $1 + 1 + 1 = 12$, for example, depending on which combination of case factors is present. This is to say nothing of the fact that factors affecting the work for a case may be completely external to a case, such as preferred post-MGH facility capacity constraints, and a reliable estimate of the total workload required for a case is of limited value for predicting the work required for a case on a given day.

It is possible to identify in the literature promising, and informative, avenues of inquiry even in settings that differ significantly from the tertiary hospital setting of inpatient case management at MGH[37][135][91]. Baillan et al. conducted their work in the context of a community mental health team responsible for elderly patients[37]. Their work focused on developing a case-weighting scale (CWS) revolving around eight case factors. The nature of this work bears similarities to the work described for Craig and Huber above, though the case factors considered differ. This is only natural given the specificity of the setting considered. While the regression-based model in the work can more aptly be described as explanatory rather than predictive, the factors identified "account" for 58% of the variability in time spent on cases in this setting (using a conventional interpretation for R^2, the coefficient of determination). The sample size, 186 cases, is relatively small, but the work is significant because it moves beyond an acuity score to consider workload quantification in terms of the time required for a case. While the setting differs significantly from that of inpatient case management at MGH, the attempt to quantify work on a scale calibrated to case manager time required and the use of a regression-based model is a significant contribution to the literature.

Based on our review of the literature, the efforts to develop a system to quantify case manager workload in a tertiary setting like MGH can best be described as nascent. Mawn provides a relatively comprehensive overview of the trends in case management acuity determination[95]. Her conclusions are telling:

> "...there are no set guidelines for all CM settings in relation to acuity and levels of intervention. Thus each CM site needs to develop clear goals and objectives in order to

determine best practices."

This conclusion is a consequence of the fact that, since there is no one model of case management clearly superior to other models, a wide array of models are currently implemented[123]. This heterogeneity precludes, to a degree, application of any existing workload quantification systems to the specific context provided by MGH.

For our work the implications of this conclusion were clear. We would have to construct an MGH-specific framework and associated analytical techniques to develop a system that allowed us to quantify case manager workload and predict case manager workload on any given day.

3.2 Text analytics applied to unstructured (free-text) healthcare data

As described in Chapter 4, the primary data source used in the analysis at the heart of this work consists of free-text case manager notes. Unstructured data of this sort presents many challenges. The primary challenge is in providing an appropriate structure for this data. In this context, appropriate can be interpreted to describe a structure that allows the use of statistical or machine learning techniques to operate on the data.

Much of the early work employing free-text data focused on text categorization to answer questions of the form "is this body of text similar to that body of text?" The basis for this work was a bag-of-words (BOW) representation of text[40][71][120][68][97]. In a BOW model a sample of text is reduced to a count of constituent words; much of the semantic information, including word order and underlying grammar is lost. Despite the simplicity of the BOW technique, and in many cases because of this simplicity, it has much to recommend it. In many applications, such as document classification or sentiment analysis, the frequentist approach implicit in a BOW model is not a significant limitation[94][130][107]. To explain, in the most general terms, we can consider the primary limitation of a BOW approach as stemming from a loss of relevant context. However, the organizing topic for textual analysis is often known, be the topic a Twitter hashtag or a specific product in a customer support call canter setting[86][106][30][102][92]. In these cases the relevant context remains intact. A careful analysis of more sophisticated natural language processing techniques reveals an underlying dependence on a word count/word frequency approach.

Much of the data that exists in our world, regardless of domain, is in the form of unstructured, free-text data. This is certainly true in the healthcare domain. Many attempts to wrangle free-text healthcare data are documented in the literature; efforts of this sort are also in progress at MGH. Given the importance of text analytical techniques for our project, it is instructive to review recent text-analytical efforts in the healthcare domain.

Temple et al. describe the use of text-analytical techniques to identify patients in a neonatal intensive care unit (NICU) that will be ready for discharge in the ensuing 2-10 days[122]. This work conducted a retrospective examination of progress notes for 4,693 patients requiring 103,206 patient-days of hospitalization. The study employed a BOW approach, using single words and bi-grams(groups of two words), to identify patient cohorts that performed poorly in the NICU based on progress note text. As the study authors freely admit, their work is preliminary and would

benefit from refinements they propose. In its current state the authors conservatively estimate that their BOW-informed supervised machine learning algorithm (random forest classifier) could potentially avoid over 900 days of hospitalization by identifying cohorts that should be targeted for more intensive discharge planning intervention to, for example, arrange for post-discharge children services, procure home medical equipment, or provide requisite parental education. As a notable aside, this work identified social issues, in contrast to purely clinical issues, as a major factor in delayed discharges from the NICU.

Temple et al. use text analytical / text mining techniques to identify cohorts of NICU patients distinguished, in part, by the amount of intervention required to effect a timely discharge from the NICU. This concept of cohorts is flexible and suggests that text analysis could be used to identify, retrospectively, cohorts of patients distinguished by any number of characteristics including the amount of work required over the course of a patient's LOS, or by the amount of work required on a specific day during a patient's LOS.

While Temple et al.'s work was specific to the NICU setting, it builds upon earlier work that provides support for the above assertion that text-analytical techniques can be used to identify patients, based on case note text, with a wide range of characteristics of interest. Yang et al. detail a hybrid text-mining approach, employing dictionary look-up, rule-based, and machine learning techniques, designed to identify the presence of a disease in clinical discharge summaries[134][133]. This method achieved a micro-averaged F-measure of 0.81 in detecting the presence of target diseases explicitly stated in discharge summaries. The method also achieved a macro-averaged F-measure of 0.66 when identifying the presence of diseases not explicitly stated in discharge summaries.

Wright et al. describe the use of a support vector machine (SVM) to categorize free-text notes forming part of the electronic health record (EHR) for patients[131]. Specifically, the SVM was trained to identify case notes indicating a patient had diabetes. This method achieved an AUC and F-measure of 0.947 and 0.935 respectively. In addition to supporting the utility of a text-mining approach in general within a healthcare context, these results suggest that text-mining could be used to support a case manager's efforts in automating some of the work required to complete a high-risk screen or initial assessment. Though formally beyond the scope of this thesis, the potential benefits of such automation are briefly discussed in Chapters 6 and 7.

In contrast to the paucity of literature related to case manager workload quantification in a tertiary setting similar to MGH, examples of the potential applications of text-mining free-text notes for a patient abound in the literature. In addition to being a worthy contribution to the literature in its own right, Temple et al.'s recent work[122] provides an extensive bibliography revealing the use of text analytical techniques to identify (or extract information from) patients with asthma[132], celiac disease[93], and epilepsy[61][58], as well as identifying patients who developed pneumonia at some point during their hospitalization[39].

Text-analytical techniques, used in conjunction with machine learning techniques, play a key role in our work as we develop a way to automatically score the amount of work required for a case. Identifying the presence or absence of a disease or other feature based on text notes is fundamentally different, and in most respects easier, than using text analytical techniques to score the amount of work documented in free-text notes. Yet, as described in Chapter 4, an augmented BOW approach forms the core of the text-analytical engine we use to develop the work score for a case. This work score, following validation, is then available for use as a response variable in the explanatory and

predictive models we develop.

3.3 Machine-learning techniques for imbalanced data sets

Many machine learning algorithms were examined over the course of our work. Though this thesis presents performance results for regression trees (when scoring the work completed for a case), linear regression-based predictive models (when predicting the workload faced by a case manager during the coming day), and boosted classification trees (when predicting whether the coming day will be a high, medium, or low workload day for a case manager), the choice of the specific machine learning algorithm employed was usually a secondary consideration. In most cases a neural net, SVM, logistic regression, knn classifier, or random forest, for example, could have been employed with similar performance results. The choice of specific algorithm was usually driven by either a desire to develop the most transparent model possible, without sacrificing performance, or by the structure of the data naturally resulting from any pre-processing techniques or analysis frameworks we developed.

More important than the specific algorithm used were the techniques employed to train several of our models. Most of these techniques were necessitated because some of the sub-problem specifications detailed in this thesis resulted in imbalanced data sets. This was most evident in our early work to predict patients (cases) that would require a high amount of work from case managers over the course of their LOS. This class imbalance was also a salient feature when predictively classifying an upcoming day for a case manager as high, medium, or low workload.

Class imbalance can be defined generally, from the perspective of machine learning classification algorithms, as when there is a clear minority and majority class. This definition lacks rigor but, as discussed by Weiss, most practitioners have at least an implicit categorization for the degree of class imbalance[128]. For example, Weiss lists guidelines for qualitatively categorizing the amount of imbalance: 10:1 = modestly imbalanced, while imbalance ratios exceeding 1000:1 can be considered extremely unbalanced. Yet, as demonstrated in the work of Weiss and Provost, even small class imbalances can present major difficulties for classifiers[129]. That is, even with "modest" levels of class imbalance (e.g., 2:1) classifier performance on minority classes can suffer significantly. In our work we frequently considered classifier performance in terms of identifying the top decile of patients or top decile of days by work score. Therefore, the ramifications of class imbalance informed many of the machine learning techniques we employed in this work.

In addition to providing an exposition on the problem of class imbalance in the context of machine learning, Weiss also identifies three main categories of issues that should be considered when dealing with imbalanced data sets, as well as approaches that have proven useful in dealing with these issues[128] . The three main issue categories, as identified by Weiss, are:

1. Problem definition issues
2. Data issues
3. Algorithm issues

In our work we utilized techniques to address all of these issues; a discussion is provided in Chapter

6 and in Appendix A.

Considering first data issues, we utilized specific sampling methods discussed in the literature to partially address issues arising from class imbalance. Though the utility of simple oversampling or undersampling is subject to question and results are equivocal when dealing with an imbalanced learning problem, other methods employ a more refined approach to sampling[80]. Specifically, we made extensive use of the synthetic minority class oversampling technique (SMOTE)[54]. This technique was used on training sets for classifiers to construct new, synthetic examples of the minority class (e.g. high workload patients or high workload days) by randomly assigning characteristics from the five nearest neighbors of an actual minority class exemplar in a training set. By expanding the decision boundary for the minority class in our feature space, the SMOTE method allowed us to sidestep some of the problems of overfitting that simple oversampling of the minority class, for example, may introduce[54].

The SMOTE technique allowed us to alter the bias of our classifier. From one perspective, using the SMOTE technique allows training of a classifier with a more general bias[54]. Yet from an equally valid perspective this increased general bias results from constructing a classifier more biased to the minority class. Thus, the problem of overfitting, while lessened, is still a major concern when using any sampling-based technique, no matter how refined, to address class imbalanced learning. However, at the algorithm-level a literature review reveals techniques that preserve the benefits of an intelligent sampling technique like SMOTE while decreasing undesirable consequences. In the context of our work the AdaBoost technique[72] and AdaCost technique[67] (algorithmic techniques) when used to train a classifier with a SMOTE-augmented training set, stand out as very relevant. AdaCost, by modifying the AdaBoost weight update-rule assigning higher weights to misclassified cases in subsequent learning iterations, allows one to implicitly tune the costs of misclassification[67]. This tuning can easily be made more explicit through the specification of a cost matrix when using certain techniques such as Quinlan's C5.0 (or C4.5) algorithm [109].

Often the distinction between methods addressing data issues and algorithm issues with imbalanced data is not stark. Stated another way, there are techniques discussed in the machine learning literature that allow one faced with an imbalanced data set to address data issues and algorithm issues essentially simultaneously. For example, the SMOTE technique can be used in conjunction with the one-sided selection technique, another data level technique, as well as boosting techniques, operating at the algorithm level, in a hybrid approach[54][87]. In our version of one-sided selection Tomek-link concepts and a condensed nearest neighbor (CNN) technique were used on a SMOTE-augmented training set[125][77]. Tomek-links are used to remove "noisy" majority class samples. "Noisy" refers to majority class samples that, in the specified feature space and with specified feature weights, appear to have more in common with minority class samples than other majority class samples. While "noisy" is one way to categorize these majority class samples it is just as likely, in our opinion, that either some key variable is missing from the feature space (omitted variables) or the feature space dimensions are not properly weighted. While Tomek-links can be used to increase the distance between minority class and majority class samples, resulting in a better-defined decision/classification boundary, the CNN approach effectively removes majority class samples far from the decision/classification boundary. A classifier can then be trained using, for example, a SMOTEBoost algorithm designed to lessen any overfitting problems that may result from "conventional" boosting approaches[55].

In fact, we employ all of the above techniques over the course of the work described in this thesis.

However, the most effective techniques we employed related to problem definition issues. These types of issues are highly domain specific and the available literature provides scant guidance on how to attack class imbalanced data at the problem definition level. In fact, Weiss asserts that, in addition to domain-specificity, another reason why a problem definition approach to imbalanced learning has not received more attention in the literature is that this is a not "research-oriented" approach [128]. Given that we employed the techniques described above it would be disingenuous, if not intellectually dishonest, to suggest that data and algorithm level approaches are not effective. Yet, the problem definition level is fundamental and domain-specificity is not incompatible with "research-oriented". It is possible to use sophisticated machine learning techniques described in the literature as an inferior substitute for proper problem formulation. In the context of our work, problem specification is most closely tied to Chapter 5 where we describe the analysis framework that is key to our work, a framework almost inseparable from the problem definition we developed. This framework/problem specification allowed us to move past an approach based on models and systems relying on specific case factors as described in the first subsection of this chapter.

To close out our literature review discussion it is important to note that performance metrics suitable for essentially balanced data sets, such as precision, recall, or accuracy, are not suitable for imbalanced data sets. For a trivial, contrived example, consider a 9:1 imbalanced data set (majority class: minority class) of the type we typically encountered during the course of our work. Achieving 90% accuracy requires no more than classifying every sample as a member of the majority class (e.g., low workload case or low workload day). For imbalanced data sets other metrics, discussed in the literature are more appropriate. These include balanced accuracy, g-mean, area under the ROC curve (AUC), and F-measures[80]. We use all of these measures to evaluate our models, as well as other measures such as the two-class misclassification rate described in Chapter 6. Ultimately, the best measure for the performance of our models would incorporate a non-abstracted ("real-world") cost for misclassification. The available literature does not offer much guidance for the problems investigated in this thesis. However, it is conceivable that future work could ascribe a cost to misclassification in terms of dollars, case manager time, or additional days spent in the hospital.

Chapter 4

Developing a Metric to Quantify Case Manager Workload

In an 1883 lecture the Irish physicist and mathematician William Thomson, more commonly known as Lord Kelvin, stated:

> ...I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be[3].

The sentiment pervading Lord Kelvin's quote essentially guided the work described in this chapter: the fundamental idea was to develop a way to quantify case manager workload. This began with meticulous manual scoring of cases to produce a data set could be used to train and test automated scoring techniques.

As discussed in Chapter 3, there have been systems developed that attempt to quantify case manager workload, but the aim of our work was fundamentally different. To explain, existing measurement systems, though grounded in quantitative analyses, tend to lose this grounding at an aggregate level. For example, consider a patient scoring system that ranks patients on a scale from 1-5, with 1 representing a low-level of required case manager intervention and 5 representing a high-level of intervention. Certainly, within a strata of patients the actual workload differs; e.g., not all "1's" are created equal. Of course, in the aggregate this is not a significant drawback as with any validated system all correctly scored "1's" would tend to cluster rather tightly around some value of required work. Yet, how does one compare the work for a "1" versus a "5"? Does the "5" require five times as much work as a "1"? Is the aggregate work of two "2's" equal to the work of a "4"?

Overcoming limitations of the type embodied in these questions was a necessary condition for developing an improved workload metric and scoring system. Specifically, the aim was to develop a validated metric in which patient differences in the metric corresponded more closely to differences in work required from a case manager, at least above an implicit baseline level of work required for all cases. Stated more concretely with a specific example, the aim was to create a metric where a

case scoring a 20 took twice as much work as a case scoring a 10.

Developing a metric was not an end in itself but a means to allow modeling and prediction of case manager workload at the patient and, more importantly for practical implementations, at the day level. In this sense, Chapter 4 is foundational for subsequent chapters. The metric allows measurement that affords the ability to quantify the magnitude and variability in case manager workload at all scales (patient, floor, and day). It is the measurement made possible by the metric, likely after subsequent refinements, that allows, in turn, a meaningful description of the current state, modelling, developing improvement initiatives and interventions, and assessment of the effects of any implemented system improvements.

The basic approach to develop a workload metric is shown in Figure 4.1. Each of these steps is described in detail within the current chapter. The case manager notes for 1621 general/internal medicine cases were analyzed to develop the metric. 1229 of these patients were present on White 8 from 1 October 2014 – 30 June 2015 while the remainder, 392 patients, were present on White 9 from 1 April 2015 – 30 June 2015. White 8 consists of 26 beds, 24 of which are the responsibility of one case manager and two of which are the responsibility of another case manager (CM positions 38 and 39 in Appendix E). White 9 has 25 beds, all of which are the responsibility of one CM (CM position 40 in Appendix E), though the CM covering these beds varies on a daily basis to a much greater extent than on White 8. White 9 was eventually converted into a permanent "float" position, though full conversion occurred after the period under consideration.

In terms of the current state, Chapter 4 primarily considers the aggregate workload for a patient, as well as the distribution of this aggregate workload at the patient (case) level. Chapter 5 continues consideration of the current state by looking at both how the aggregate workload for a patient is distributed during a patient's LOS and how a case manager's aggregate workload (sum of daily work for all cases) varies from day-to-day. The penultimate chapter, Chapter 6, presents a model for predicting a case manager's daily workload. Finally, Chapter 7 summarizes key findings and presents recommendations for future work and, more tentatively, interventions based on the current state of the work.

4.1 Verifying minimum suitability of notes as a high-level indicator of case manager workload at the individual level

The methodology in this chapter was built on an underlying assumption that the notes for a given case provide a reasonably complete record for the work completed by a case manage. That is, the notes do not necessarily represent comprehensive documentation of a CM's workload because it was readily acknowledged by CM leadership that the notes did not document all the work completed for a patient or, equivalently, there are notes missing from the record. The implications of "missing" notes for developing a workload metric are considered further in 4.2. However, given that notes are missing it is necessary to verify the "reasonably complete" assumption before proceeding further.

Even before examining the text of specific notes, two themes common across all case manager interviews provide a key for determining the minimum suitability of CM notes as a coarse, high-level proxy for CM workload at the patient and day level. Specifically, all CMs interviewed indicated that:

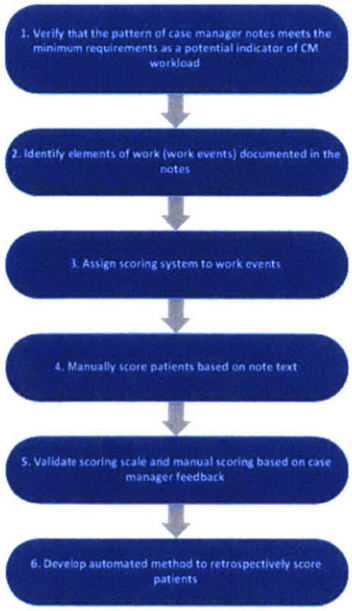
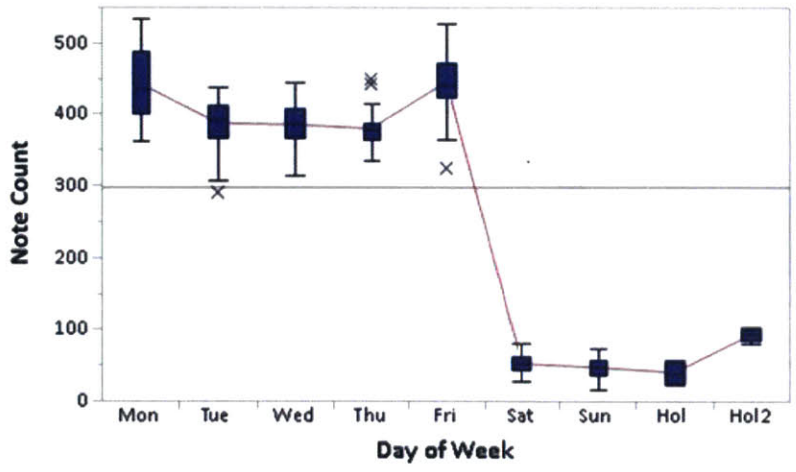


Figure 4-1: Developing a metric for aggregate workload at the individual case level

1. Friday's and Monday's are the highest workload days for CMs
2. Workload varies greatly between cases with a relatively small number of cases consuming a large amount of CM time/effort[6][8][7][1]

The boxplots in Figures 4-2 and 4-3 indicate that even a simple note count demonstrates the first global observation seems to hold across multiple perspectives and temporal scales.



	n	median	mean	sd
Mon	52	435	442	48
Tue	46	382	387	31
Wed	53	388	385	27
Thu	48	379	380	25
Fri	52	447	445	37
Sat	52	51	51	12
Sun	52	46	47	13
Hol	4	38	40	16
Hol2	5	95	93	9

Figure 4-2: Hospital-wide case manager note count by day of week, 1 Oct 2014 – 30 Sep 2015

In Figure 4-2 Hol1 refers to holidays such as Christmas, Thanksgiving, and Memorial Day, while Hol2 refers to holidays such as Columbus Day, Presidents Day, and Martin Luther King, Jr. Day. Both types of holidays result in reduced staffing levels throughout the hospital, including case managers. However, in the latter case, while in general the hospital may be at reduced staffing, for

White 8 and White 9 Hol1's have zero CM notes in the record while Hol2's have CM notes in the record (although greatly reduced to levels comparable with the weekend).

These holidays are not shown in Figure 4-3, nor are the weekends. For both figures certain days were relabeled to more accurately reflect note distribution across days. For example, a Monday falling on a holiday was not considered a Monday. Rather, the next day, Tuesday, was considered as a Monday. The histograms are shown with connected means indicated by the red lines.

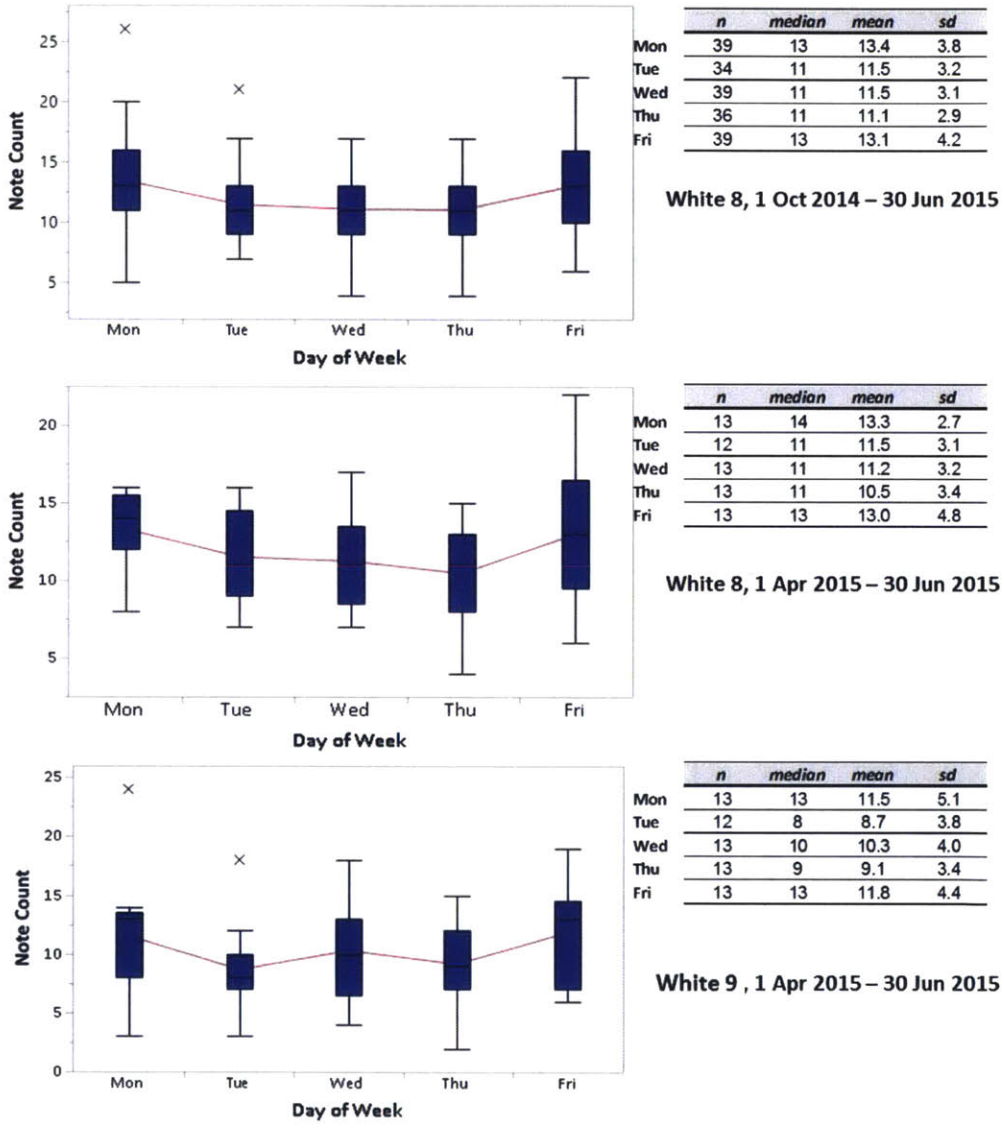


Figure 4-3: Weekday case manager note count by day of week for White 8 and White 9 over selected time periods

The weekly periodicity of CM workload, as proxied by a daily note count is clearly suggested in Figures 4-2.and 4-3. For the yearly note count in Figure 4-2, as well as the nine-month note count in the first panel of Figure 4-3, a pairwise comparison of means using the nonparametric Wilcoxon method reveals that the means for Friday and Monday are statistically different from the means of Tuesday, Wednesday, and Thursday (at a significance level below .05), while the means of Tuesday,

Wednesday, and Thursday are not statistically distinguishable. Pairwise Wilcoxon tests on the bottom two panels fail to reveal a statistically significant difference with an $\alpha = 0.05$ significance level[1]. In part, this is likely due to the small sample size, particularly for panel 2. Panel 3 is a similarly small sample, though results suggest that the periodicity may not be as pronounced for White 9. In fact, there are reasons to believe that this is true, related to the distribution of workload and work patterns across a patient's length of stay, distributions and patterns that differ between White 8 and White 9. This point is considered further in Chapter 5. The most likely reasons for the observed periodicity were presented in Chapter 2; namely, minimal staffing levels on the weekend cause discharge window work to be advanced from the weekend to Friday, while some discharge window work and admit window work is delayed from the weekend until Monday.

Considering the second global observation, inter-patient variability in CM workload, Figure 4-4 illustrates that this pattern is observed at a note count level on White 8 and White 9.

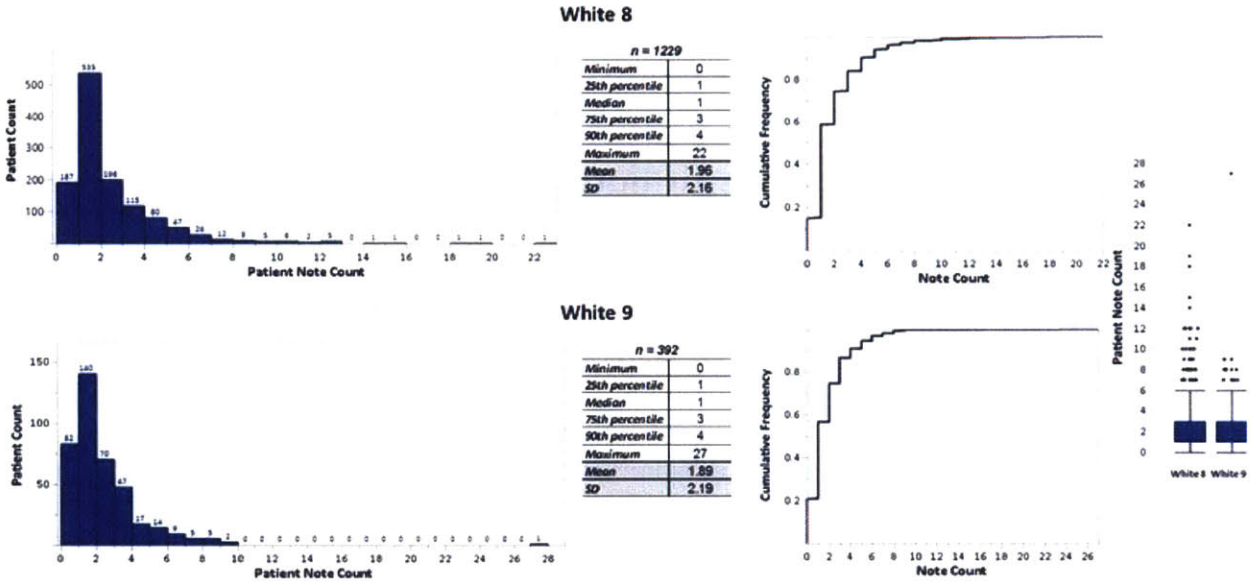


Figure 4-4: Patient note count distribution

At this point, and even given the coarseness of note count as a workload proxy, several aspects of Figure 4-4 are worth noting. First, the distribution of note count is very similar between White 8 and White 9, with identical quartile and 90th percentile cutoffs, as well as similar means and standard deviations. This is encouraging when considering extending a model developed for a small number of floors to other floors, at least floors of a similar type. Second, there are many patients/cases for which there are no notes and, hence, no work recorded. Third, the distributions of patient note count are significantly right-skewed, jibing with reports that a small number of patients require significant amounts of work. These last two aspects evident in Figure 4-4 virtually

[1]The possible issues with using multiple pairwise comparisons to draw statistical inferences have been discussed and debated for more than six decades[127][117]. In a grossly simplified form, the argument for adjustments makes use of the fact that if enough pairwise comparisons are made, even at a small value of αa result of apparent statistical significance becomes more likely. Methods to address potential issues are well-known[114], although others contend adjustments to common methods are often(or always) unneeded[112][121]. These issues are not particularly germane to the main results of our work, but all results presented for pairwise comparisons do include recommended nonparametric adjustments for multiple comparisons. A good overview of the topic is provided by Day and Quinn[63]. Rather than worrying about the more philosophical issues of adjustments, we simply made them.

guarantee, even without considering the current state at a finer level, variability in CM workload despite an essentially invariant number of cases on a daily basis (unchanging number of beds that a CM position is responsible for). Finally, the relative number of cases with zero notes seems to be higher for White 9. Much like the less stark weekly periodicity for White 9 vis-à-vis White 8 (panels 2 and 3 of Figure 4-3) it is plausible to attribute this fact, at least partially, to a difference in work patterns between the two floors. Again, this is taken up in Chapter 5.

The upshot of this section is that even a high-level indicator like note count allows one to draw insights into the current state of CM workload on White 8 and White 9. In fact, preliminary investigation of note count and correspondence with reports of CMs concerning distribution of workload by day of week and between patients suggests it may be possible, rather than using patient note count as a proxy for workload, to use note count as a regressor for a linear model given a sample of manually scored patients. This model is initially presented in section 4.5 because both the strengths and, more importantly, weaknesses of such a model suggest the model that is ultimately presented in section 4.5, as well as diagnostics that must be completed to fully evaluate model performance and limitations.

4.2 Identifying work events in note text and developing an event scoring system

Ultimately, the patient workload metric and derived metrics are intended for use as response variables to facilitate modeling and prediction of case manager workload. Therefore, the first stage of metric construction involved reading a total of 3147 CM notes, covering 1351 of the 1621 patients for which notes exist in the record, to determine what type of work events are identifiable in the notes. Depending on how one categorizes events, the list of work events could be quite substantial. Consider the work event of a phone call. This work event category could be further divided, *ad nauseum*, into an array of subcategories, such as phone calls to family, phone calls to post-MGH subacute facilities, phone calls to durable medical equipment providers, and so forth. The problem with such a fine division of workload events is that it becomes increasingly difficult to identify events with this degree of specificity when using text analytical techniques, a cornerstone of the automated method presented in 4.5. Instead, scoring focused on the limited set of workload events shown in Table 4.1[2].

Scoring is done from the perspective of the DCP CM on White 8 or White 9. There is an imputed time value for each unit of score derived from the low-end value assigned to the high risk screen and the high risk screen plus initial assessment, 2 and 4 respectively. Since reports and observations indicate that the HR screen takes 10 minutes to complete in its current incarnation, while the HRIA takes 20-25 minutes, each unit of score represents approximately 4-6 minutes. Similarly, faxing, assigned a score range of 2-3 has an imputed time value of 10-15 minutes. These imputed time values are very tentative (and tenuous) which is why we use the unitless score, rather than time values directly, for analysis. We were only able to record directly the time associated with 67 work events during the later observation periods (which did not occur on White 8 nor White 9), as a protracted time-motion study was not in the scope of our work. Furthermore, the time values

[2]It was stated explicitly in Chapter 2, but it bears repeating, that distilling a case manager's workload into the limited set of events in Table 4.1 should by no means be construed as downplaying the potential complexity of the work for a given case.

Table 4.1: Work event scores in case manager notes

Workload Event	Scoring Range
High risk assessment	2-4
High risk and initial assessment	4-9
Meeting with patient or patient's family	2-9
Phone call	1-7
Placing referral	1
Email	1-2
Faxing	2-3
Coordination with care team member	1-4
Reviewing chart	2-4
Documentation	1-2 (per note)*
Arranging transportation	1

* Other forms of documentation may include face-to-face forms for VNA long-term care forms, MOLST, or DNR forms.

for events that were recorded are not matched with the accompanying documentation; cases span multiple days (or weeks) and it was rarely possible to observe the work for a case from start to finish. Finally, there are only a handful of events (23) for which a time value is indicated in the notes. The values of event scores in relation to each other were deemed "not unreasonable" by two case managers who assisted with the validation of the case scoring presented in Section 4.4. Admittedly, "not unreasonable" is not a glowing endorsement of the scoring system, but the validated results presented in Section 4.4 suggest that the scoring system is is useful even in current form.

The scoring system also reflects the fact that HRIAs, meetings (with patients, families, or other members of the care team), and phone calls (of various types), as well as documenting these events, constitutes the bulk of CM work. The range of scores given for certain events is indicative of the fact that, for example, not all phone calls represent an equal workload. In manually scoring events with a range of possible scores, the duration of the event was used to assign a single score value. The information exchanged during the event and/or the tenor of the event was used as the main indicator for duration. Continuing with the phone call example, a brief phone call to a facility to ascertain whether any beds are available may be scored as 1-2, while a phone call to a patient's family member to discuss discharge planning options may be scored as a 4-5. Similarly, a meeting or exchange to let a patient know that she has been accepted to her first choice post-MGH facility may be scored as a 1-2, while a family-team meeting to discuss discharge planning options for a terminally-ill patient with limited financial means and an overwhelmed and/or non-cooperative family could be scored as a 6-9. In general, when manually scoring cases the relevant length of a note (i.e., the length of the text detailing a work event and not, for example, clinical information about the patient) was used to score events for which a range of values was possible.

The scoring system in Table 4.1 is rather naïve and the system, as well as the utility of a constructed workload metric response variable, could benefit from further refinement based on time-motion studies of an extended duration and across multiple floors. Some more salient weaknesses and omissions bear noting. One omission is the work done for a patient after a discharge. To explain, case managers were observed performing their work over a period of 16 sessions totaling in excess of 81 hours. In 7 of these 16 sessions a case manager performed work related to a patient that had already been discharged, usually prompted by a phone call from a patient or a patient's family. The

nature of this work varied and it was, more often than not, due to either some oversight by another member of the care team (such as neglecting to send documentation or imagery to the post-MGH facility), or in trying to resolve a complaint about post-MGH circumstances. In two cases a DCP CM was observed spending significant time to answer a patient query about billing issues. However, despite the prevalence of this type of work during observation periods, only six instances of this work show up in the notes examined.

Another shortcoming is in accurately scoring the various documentation that a case manager may have a hand in completing for any given case[3]. For example, the long-term care (LTC) form that must be completed for patients with MassHealth as the primary insurance takes 90 minutes to complete. Completing this form alone would result in a score of 12-16. For this particular form we had access to a report listing the number of LTC forms completed by floor for each quarter and no forms were listed during the time period encompassing the cases examined. However, there are other forms that a CM may have a role in completing underscored for some cases.

Even more problematic, particularly for automated scoring, is that the forms could be the responsibility of the CM or another member of the care team, depending on the floor. This extends to other forms of work as well. For example, on White 8 the CM was less involved in coordinating the resumption of previously established home care services (services that existed before the current hospital admission), though usually a primary facilitator when coordinating for new home services; coordinating resumption of services was usually the purview of another nurse. In contrast, the White 9 case manager was typically involved in coordination for resumption of home care services. Similarly, on some floors the assigned CMRS may assume a greater role than on other floors. Differential role boundaries are easily accounted for in manual scoring, but are an obstacle for automated scoring and, especially, extending an automated scoring method to other floors. This obstacle and ways to limit the effect on automated scoring are considered in section 4.5.

In addition to differential role boundaries, shifting role boundaries are also evident upon a reading of the notes; i.e. when a CM performs a task in one case and another member of the care team performs the same task for another case. This is not an issue when considering the sum of work completed for a patient regardless of who completes it, but the scoring system is based on the amount of work specifically completed for a case by the White 8 or White 9 CM. Many CM notes, in addition to documenting work completed by the CM (direct action) also detail work completed by other members of the care team (referred actions). The presence of both shifting role boundaries and referred actions, while controlled for in manual scoring, can greatly complicate the task of automated scoring (see section 4.5). Of course, one can raise many other objections to the scoring system, such as potentially under-scoring the amount of work done by CMs in reviewing clinical documentation.

The potential deficits for an individual case score were enumerated above. There are also elements of work not captured when aggregating case scores. For example, the amount of time that a case manager spends in daily rounds would not be captured by summing all of the work documented in CM notes for a given day. That is, there is a baseline amount of work distributed across all cases completed by case managers on a daily basis; this baseline work likely varies by floor.

A note may document many individual events and, despite the relatively limited number of events and range of scores possible for events, scoring cases manually resulted in a wide range of scores

[3]All forms of documentation were scored.

across patients. The results of manual scoring are presented in the next section, while section 4.4 details the method used to validate scoring. The scoring system is the foundation upon which the workload metric is built and, again, a formal time-motion study, matching event duration with accompanying documentation, would undoubtedly lead to valuable refinement of the scoring system and assist text-based automated scoring.

4.3 Manually scoring cases

Using the system in Table 4.1, a score for each of the 1651 patients/cases examined was manually assigned. The results of this manual scoring for White 8 and White 9 are shown in Figure 4-5.

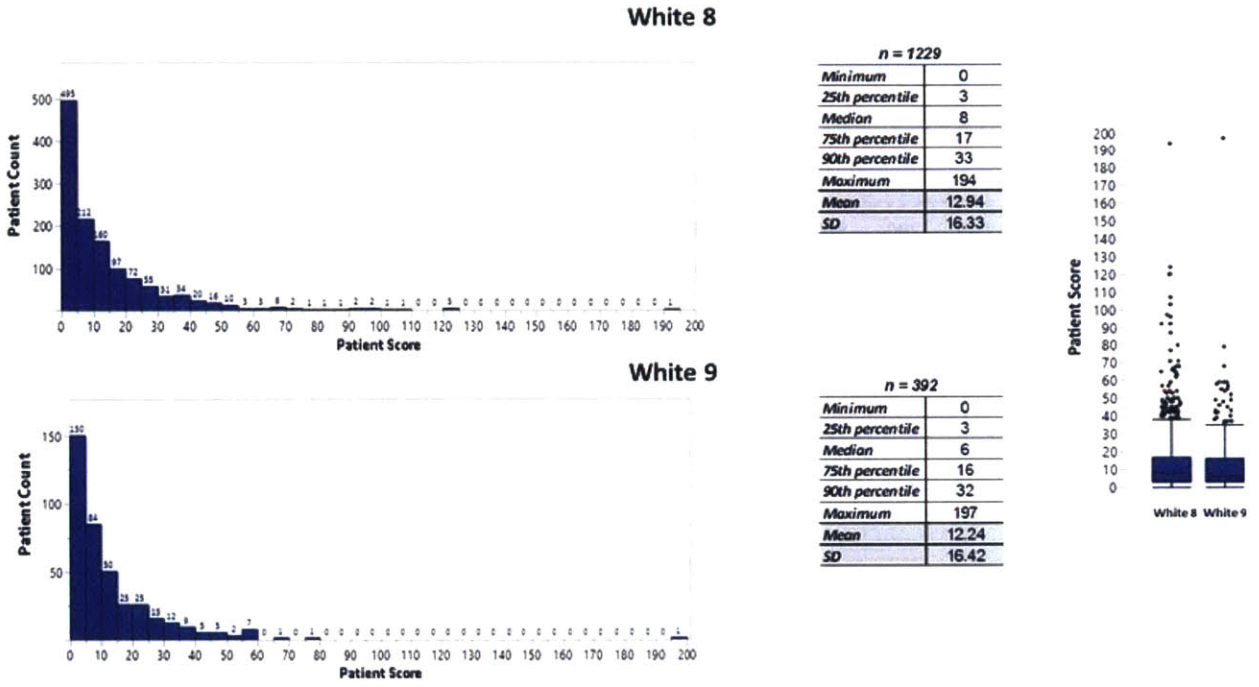


Figure 4-5: Distribution of individual patient (case) scores for White 8 and White 9

Similar to Figure 4-4, the distribution of individual patient scores on White 8 and White 9 are very similar. Furthermore, there are a relatively small number of patients that require a disproportionate amount of work from the case manager. Figures 4-6 and 4-7 illustrate this last aspect of inter-patient variability of workload in a more pointed way. In these figures the phrase "middle 30% of patients" refers to patients in the 60th – 90th percentile of score. This division into bottom 60th percentile, 60th-90th percentile, and top decile is a legacy of earlier work focused on assigning a classification to patients rather than a numerical score, but it is still useful in making the point about unequal distribution of work across patient classes. The red dashed line intersects at the top 10% work score cutoff point when patients are ordered by decreasing score on the x-axis.

On both floors the top 10% of patients manually scored account for nearly 40% of the total documented work. The middle and bottom groupings of patients similarly account for a comparable amount of work.

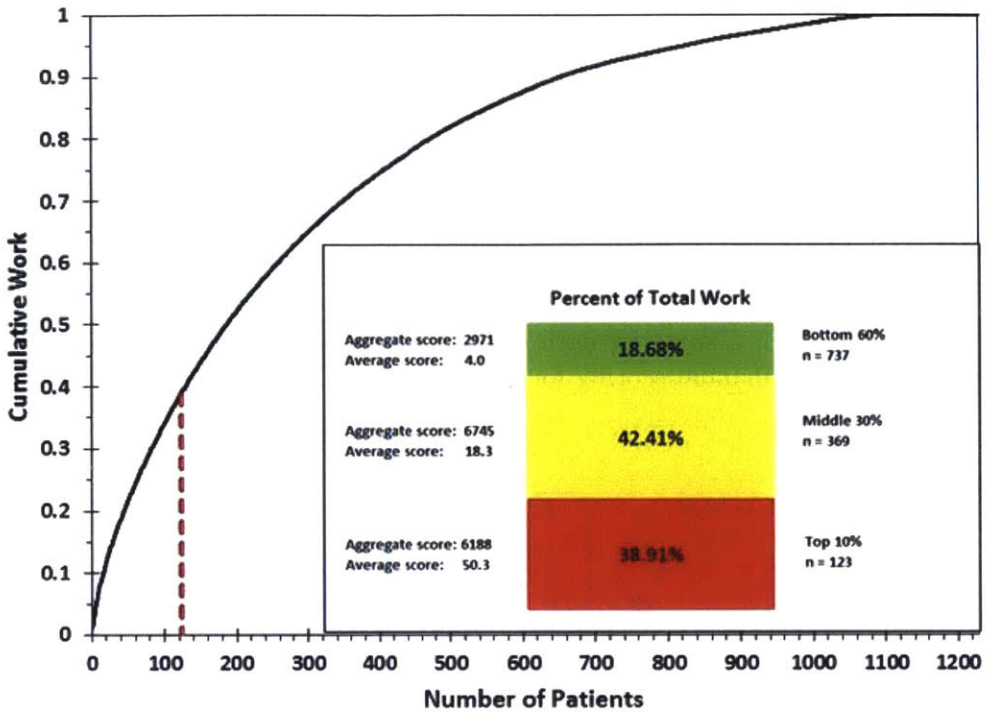


Figure 4-6: Distribution of work across patient groups for White 8, n=1229

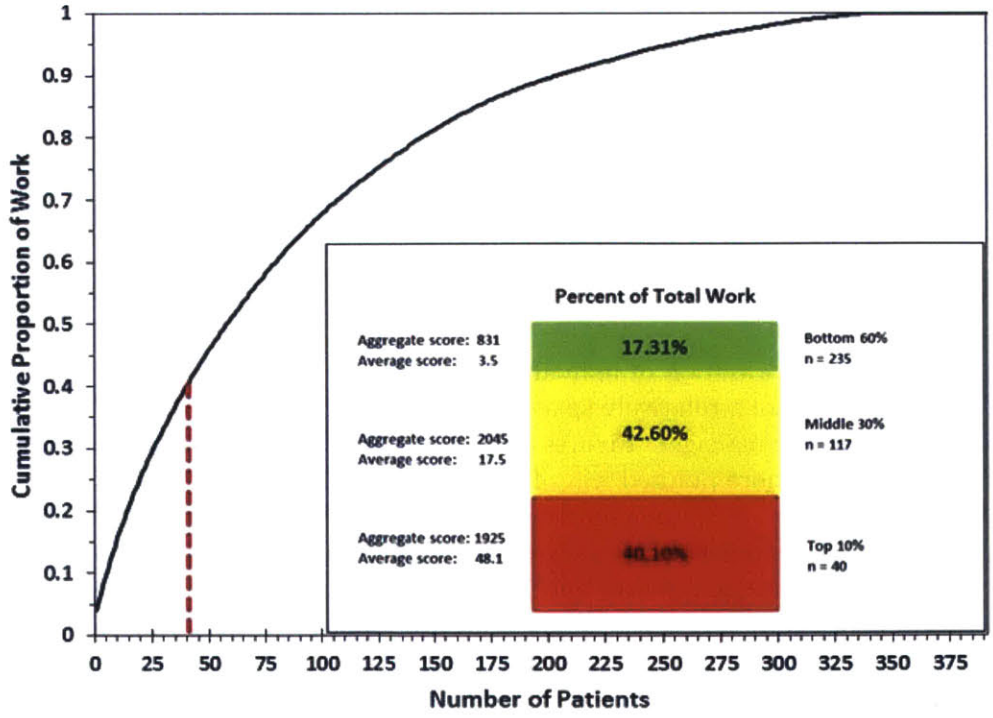


Figure 4-7: Distribution of work across patient groups for White 9, n=392

As alluded to multiple times, the work documented in CM notes for a patient almost certainly does not capture the entirety of work for a given case. However, there is an assumption that patients requiring more work will have more work documented. If this assumption holds and there is not a gross systematic under-documentation of work for patients then the scoring methodology, if validated, still retains its utility.

There is some evidence of a minor systematic under-documentation of work, not by patient type, but rather by day of the week. That is, there seems to be a greater chance of missing notes, and hence missing documented work, for cases that CMs perform work for on Fridays and Mondays. This possible systemic under-documentation is examined more fully in Chapter 5. However, it stands to reason that, given that Fridays and Mondays are typically the busiest days for CMs, documentation of work may be of lower priority for CMs as they are engaged in actually performing the work of facilitating the discharge of patients.

It is possible to estimate the "missing" work for patients, in effect counting work that is not documented in the notes, by considering either the typical sequence of notes for a patient and/or imputing missing work by relying on other sources.

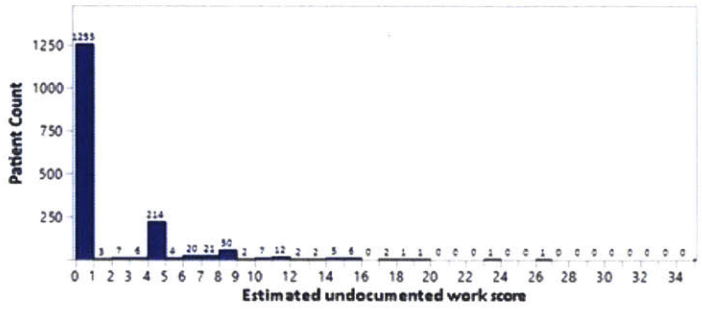
As an example of the first method for imputing missing work consider a patient admitted from home that eventually discharges to a skilled nursing facility. At a minimum this patient would usually have a HRIA note, a note documenting a meeting with the patient to get facility preferences and place referrals, and a note indicating the patient had been accepted to a facility, the patient agreed with the placement, and transportation was arranged for the patient (an inpatient facility transfer note). If the record only shows an HRIA completed before it was determined SNF placement was required and an inpatient facility transfer note, obviously the work events of a meeting to receive patient preferences and subsequent placement of referrals are missing from the record[4].

As an example of the second type of inference for missing work, a patient that is difficult to place may be denied by multiple preferred facilities. There may be a note in the record indicating a meeting where the CM informed the patient of the denial(s) and asked for additional facility preferences. If this meeting was documented as ending with the patient requesting more time to consider facility choices and the next note is an inpatient facility transfer note to a facility for which no preference or referral is documented, then an intervening meeting may have occurred. There is a referral database that can be examined for some cases. If this database indicates a referral between the dates of the meeting where the patient was informed of facility denial(s) and the transfer note then it is reasonable to assume undocumented work occurred. Of course, estimating this missing work is an inexact process, especially as it is only reasonable to impute an average value for the missing work that is inferred.

Figure 4-8 shows an estimate for the distribution of missing work. Automating the inference of missing work is difficult but, as shown in Figure 4-8, most patients have low values of missing work. The most common type of missing work is undocumented HR screens or HRIAs. This accounts for the large count of patients with an estimated value of 4 for the amount of undocumented work score. There is a database indicating whether a patient met high risk criteria and presumably the CM did work to complete this screen. Similarly, there is also a less reliable secondary source

[4] A two-note sequence is common for a patient admiited from a SNF on a bed-hold. In these cases the CM typically verifies with a patient during the HRIA that she is willing to return to the facility when ready; the HRIA is the first note. When the patient is ready for discharge an inpatient facility transfer note is entered into the record; no intervening note documenting a meeting to get placement preferences is required.

indicating that an HRIA was performed, even if undocumented. Again, the value of this missing admit window work would typically contribute 2-6 for the case work score, hence, the high count of patients estimated to be missing 4 units of work score. For other cases, particularly cases of extreme missing work, it is likely that some notes were never copied into CAS so that they were not retrieved in the data pull of CM notes provided for analysis.



n = 1621	
Minimum	0
25th percentile	0
Median	0
75th percentile	0
90th percentile	4
Maximum	35
Mean	1.33
SD	3.01

	White 8 n = 1229	White 9 n = 392
Documented Work Score	14069	4100
Estimated Missing Work Score	1590	573
Total	15659	4673

Figure 4-8: Distribution of estimated undocumented work (score) for White 8 and White 9

In effect, each patient has two scores, a score based solely on the notes and a score based on the notes plus implied work. The former of these scores is used in developing the text-analytical based method of automated scoring in Section 4.5. In contrast, the implied work score is used when developing the model to predict the aggregate daily work score for a case manager in Chapter 6. In any case, the correlation between documented and documented plus implied work is very high, with a Pearson correlation coefficient (parametric) of 0.9848 (95% CI: 0.9833,0.9862) and a Spearman correlation coefficient (nonparametric) of 0.9682. For White 8 the documented manually scored work is 89.8% of the total documented plus implied work score, while for White 9 this figure is 87.7%.

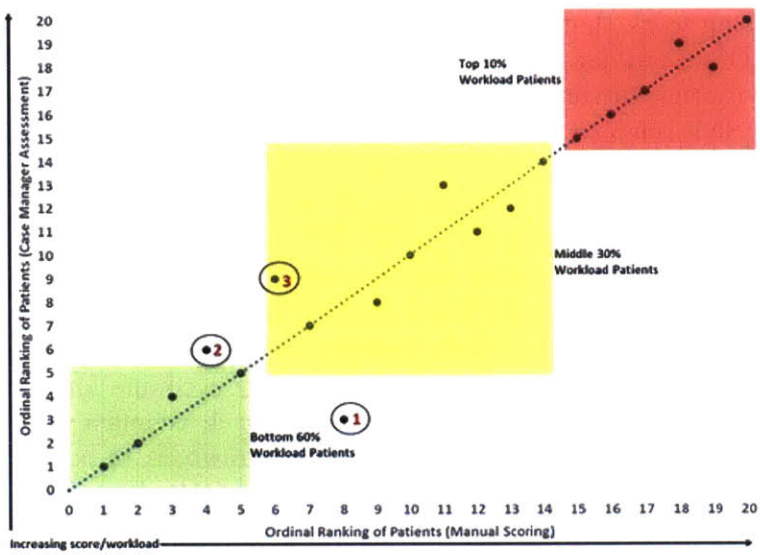
4.4 Validation of manual case scoring

As when considering the validity of note count as a high-level indicator of work in section 4.1, the results of manual scoring seem to conform to expectations based on case manager reports that a relatively small number of patients requiring a disproportionate amount of total work. Of course, reading thousands of CM notes and subsequent manual scoring is very time-consuming. Furthermore, it is possible that the method of manual scoring assigns the wrong patients the highest scores. Developing an automated method of scoring the work metric requires validation so that the metric can be used as a response variable to train a model to score the text of CM notes.

Two seasoned case managers were enlisted to help validate the scoring system; these case managers were not affiliated with the cases. CM1 was a "float" case manager that works on a number of different floors during the course of a week. CM2 is a clinical nurse specialist who is frequently assigned to help DCP CMs with very complex cases. Each case manager was given a list of 20 cases to review across the manual scoring spectrum. Rather than scoring the cases, the CM auditors were asked to independently rank their assigned cases, from 1 to 20, in order of increasing work required by case managers. The CM auditors were free to use all documentation associated with the case in making their ranking determinations, including clinical notes we did not have access to when scoring

the cases. Ten of the cases overlapped between case managers to screen for noteworthy differences in rankings between auditors. The auditor ordinal ranking of cases was then compared with an ordinal ranking determined by the manual score of the case. A comparison of ordinal rankings also allowed an inference as to whether, in a more qualitative sense, high, medium, and low workload cases were correctly scored. Matching an auditor's ordinal rank is a more rigorous task than matching a qualitative classification. The implied plus documented workload score was used in ranking patients. This method of validation is not comprehensive, but it can reveal whether higher manual scores are indicative of higher CM workload for a given case.

The results of comparing CM1's rankings with rankings derived from manual scores are shown in Figure 4-9.



Manual Scoring Ordinal Rank	Case Manager 1 Ordinal Rank	Absolute Rank Difference	Manual Case Score
1	1	0	0
2	2	0	3
3	4	1	8
4	6	2	9
5	5	0	10
6	9	3	12
7	7	0	12
8	3	5	16
9	8	1	18
10	10	0	20
11	13	2	21
12	11	1	22
13	12	1	28
14	14	0	30
15	15	0	35
16	16	0	42
17	17	0	59
18	19	1	67
19	18	1	68
20	20	0	71

Figure 4-9: CM1 ordinal ranking of 20 cases compared with ordinal ranking corresponding to manual score

The layout of Figure 4-9 may require some explanation. The scatterplot in the left panel plots the CM's ordinal ranking of cases versus the ranking that arises from a comparison of the manual case score. The dashed 45° degree line represents a line of perfect agreement; i.e., if the manual score ranking and the ordinal ranking agree then points will lie on this line. The shaded rectangular regions provide an indication of whether the rankings imply the case would have been ranked in the same groups of high, medium, and low work. The same percentile cutoffs described in section 4.3 are used for this derived classification: high work is the top decile of patients by manual score, medium work represents cases in the 60th – 90th percentile of manual score, and low work is the bottom 60% of patients by manual score. Again, the medium workload patients are described colloquially as the "middle 30% workload" patients. The classification is only implied because auditors ranked cases by increasing workload rather than formally classifying patients as high, medium, or low workload. Points lying within a shaded region have the same implied classification.

The right panel of the figure provides the manual score case ranking, CM1's ranking, the absolute rank difference, and the manual case score. Absolute rank differences greater than 2 are highlighted in red. These points are also circled and numbered on the scatterplot, as are points whose implied classification does not match. For manual case scores, low, medium, and high workload cases are

highlighted green, yellow, and red, respectively.

Even a cursory examination of Figure 4-9 reveals substantial agreement between manual score-derived ordinal rankings and CM1's ordinal rankings. 18 of 20 cases have the same implied classification. There are, however, two cases with an absolute rank difference exceeding two and these deserve discussion as to the probable reasons for the rank disparity.

Consider first the point labeled 1 in the scatterplot. This case was ranked as 8 by manual scoring and as 3 by CM1. This is a case for which there is implied missing work, work which is accounted for in the manual score, but is not explicit in any documentation. The explicit documented work would score as an 8 (with a rank of 3).

However, there is reason to assume missing work that would have been scored, on average as 8 (total score of 18). This patient had an LOS of ten days and in the record there is a weekly update note and a note documenting a meeting with the patient to discuss referrals to a SNF. The patient's ultimate discharge disposition was to the SNF referenced in the second note and transportation to the facility was via ambulance based on the raw data from a CM workload report. However, there is no note documenting the inpatient facility transfer, with a typical work score of 4-6. This work would consist of a brief meeting with the patient to inform the patient of facility acceptance and transfer time, arranging transportation, final coordination with the care team, and documentation of the transfer. There is also evidence of an undocumented HR screen for the patient in that the CM workload report indicates a HR screen was completed for the patient, with a typical score of 2-4. The difference between documented and document plus implied work likely accounts for the difference in rankings. Of note, the manual score ranking would have been 3 or 4, versus a CM1 ranking of 3, if implied work was not included in the manual score. This is one drawback to having independent auditors of cases used in validation of the scoring methodology; the CM who worked on the case would have likely accounted for the undocumented work when ranking the cases.

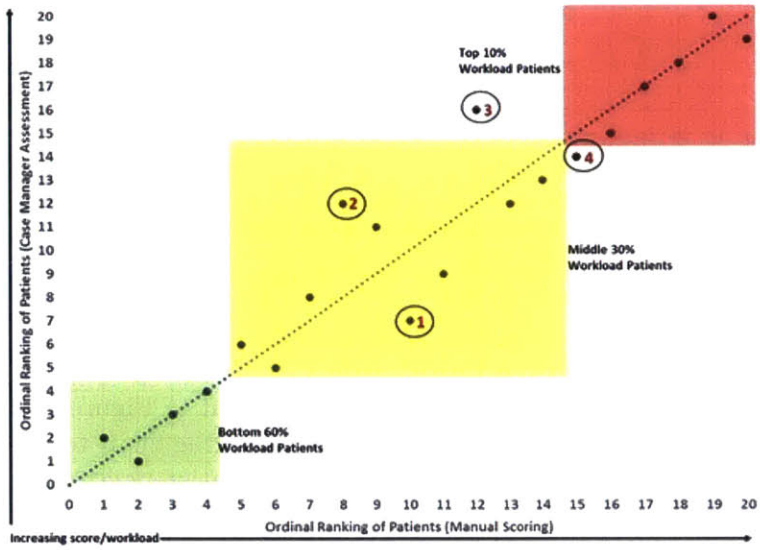
The reason for the difference in rankings for the point labeled as 3 in the scatterplot is more straightforward and relates to the boundaries used when scoring the case. As explained, patients are scored from the perspective of the CM on either White 8 or White 9. The auditors, however, scored the cases based on the amount of CM work required by all CMs who worked on the case. For this particular patient a HR screen was completed and it was determined that, though the patient met HR criteria, this patient did not require CM intervention; this work took place on a different floor and by a CM not on White 8. The patient subsequently transferred to White 8 where discharge planning work was required. In manual scoring the work completed on the other floor is not counted and this results in a higher CM1 ordinal ranking as compared to the manual score. It is necessary to use our work boundary, the work completed by the White 8 CM for the patient, rather than an expanded boundary counting all work completed by all CMs for the patient, if daily workload for a specific CM/floor is to be modeled and, ultimately, predicted in Chapter 6.

Looking at the right panel of the figure, many of the cases were scored manually at close intervals. However, two cases scoring 12, or a case scoring 21 versus 22, are actually significantly different from a component perspective. While the aggregate manual scores are comparable, the composite work events and note text leading to these scores is highly variable. The cases were selected to help ensure the scoring methodology was valid across highly varied text and work events.

There is a more subtle effect in the ordinal ranking provided by CM1 that only became clear in discussions with the auditors after the validation process. Specifically, if cases score relatively

closely in the aggregate sense then the auditors assigned the higher rank to the patient with the shorter LOS. Obviously, for a given level of aggregate work, a shorter LOS gives the CM fewer days to complete the work and the case presents a more intense workload with intensity measured as (aggregate work/LOS). As an example, consider the cases ranked 12 and 13 by manual scoring (22 and 28 respectively). The aggregate work events for these cases are very similar except the case manually ranked 13 has an extra note detailing two brief phone calls to a family member. However, the case with the manual ranking of 13 (28 work score) was for a patient with an LOS of seven days. In contrast, the patient with a manual ranking of 12 had a lower work score of 22, but the LOS was only 3 days. CM1 ranked the case with the lower score of 22 as representing more work for a CM. This effect calls into question the very notion of a "high workload" case. Is this high aggregate workload, high intensity, high peak workload, or some composite of all of these attributes? This idea is explored further in Chapter 5 and in Chapter 6.

Figure 4-10, identical in format to Figure 4-9, presents CM2's ordinal ranking of cases in comparison with the manual ranking. As with Figure 4-9 there is substantial agreement between the two sets of rankings, though there are some points of difference to be addressed.

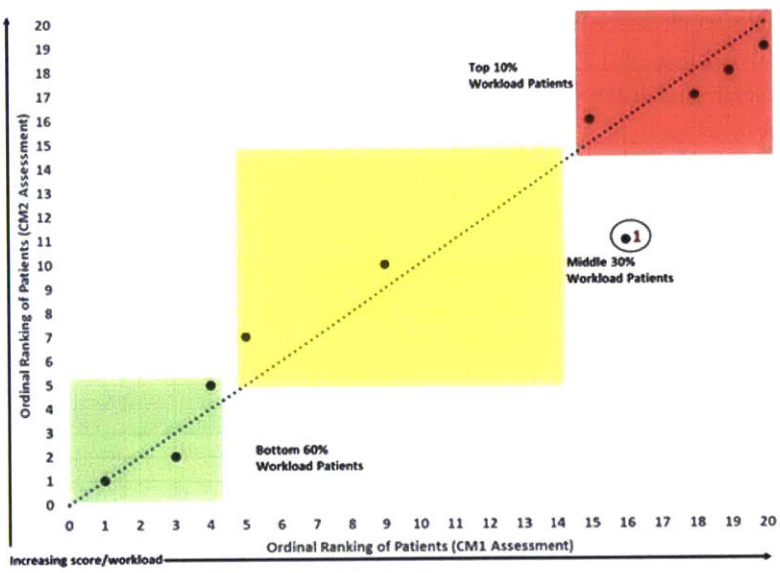


Manual Scoring Ordinal Rank	Case Manager 2 Ordinal Rank	Absolute Rank Difference	Manual Case Score
1	2	1	0
2	1	1	0
3	3	0	3
4	4	0	10
5	6	1	11
6	5	1	12
7	8	1	14
8	12	4	14
9	11	2	15
10	7	3	18
11	9	2	21
12	16	4	22
13	12	1	27
14	13	1	28
15	14	1	[illegible]
16	15	1	[illegible]
17	17	0	[illegible]
18	18	0	[illegible]
19	20	1	[illegible]
20	19	1	[illegible]

Figure 4-10: CM2 ordinal ranking of 20 cases compared with ordinal ranking corresponding to manual score

The case manually ranked 8 (CM2 rank of 12) represents another instance of a case where a patient transferred onto White 8 resulting in some of the work done by another CM on another floor. The implications of this for the two sets of rankings was discussed above. Similarly, the case manually ranked 10 with a CM rank of 7 is likely due to the inclusion of undocumented work to derive the manual rank. In this case the referral database lists an undocumented referral that would have likely required some type of communication with the patient; at the very least there would have been a brief meeting and referral placed. The rank difference for the case manually ranked 12 with a CM2 rank of 16 represents a second type of work boundary issue. This patient required an inpatient psychiatric admission, ultimately being admitted to Blake 11. As presented in Chapter 2, there is a CM assigned to provide psychiatric consult services for patients on hospital floors requiring acute inpatient psychiatric placement. The psych consult CM does not completely alleviate the work burden for other floor CMs imposed by these patients, as the floor CM must communicate

(phone calls, meetings/case discussion, emails, etc.) to help facilitate the required transfer. However, the psych consult CM does assume most of the work burden for these patients. CM2 accounted for the entirety of this work in ranking the case and, possibly overestimated the amount of work required of the White 8 CM. Figure 4-11, comparing CM1 and CM2 ordinal rankings for the 10 overlapping cases illustrates this last point.



Case Manager 1 Ordinal Rank	Case Manager 2 Ordinal Rank	Absolute Rank Difference	Manual Case Score
1	1	0	0
2	3	1	3
5	4	1	10
7	5	2	12
10	9	1	21
11	16	5	22
16	15	1	42
17	18	1	58
19	20	1	67
18	19	1	68

Figure 4-11: Comparison of CM1 and CM2 ordinal ranking for 10 cases

Point 1 in the Figure 4-11 scatterplot shows the comparison between the CM1 and CM2 ranking. Of note, CM1, the "float" CM, frequently serves as either the Blake 11 CM or the psych consult CM and has a better experience base to judge how the work for a patient requiring inpatient psychiatric admission is divided. Comparing the scatterplots in Figures 4-9 and 4-10 with that of Figure 4.11 indicates the CM1 ranking for this case is much closer to the manual score ranking than to the CM2 ranking. With this acknowledged, Figure 4-11 does indicate the desired consistency in how experienced CMs would rank the validation cases in order of increasing workload.

Similarly, the agreement between manual score rankings and CM rankings, as noted, is substantial. We can consider a variety of nonparametric measures to statistically quantify the strength of agreement among manual and CM rankings. Though typically used for more than two raters, Kendall's W (Kendall's coefficient of concordance) is most appropriate in this case[85]. Comparing manual rankings and CM1 rankings the coefficient is 0.9820; a comparison of manual rankings with CM2 yields 0.9789. Both of these values indicate significant agreement among the rankings, or more correctly, that the ratings apply essentially the same standard for rating, namely; ranking cases in order of increasing workload. Generally, values above 0.9 are considered good[118][73][5].

This coefficient can be used for hypothesis testing with:

[5]Other tests such as Kendall's rank correlation coefficient or even Spearman's rank correlation coefficient could conceivably be used for examining the agreement between manual score-derived rankings and CM rankings. In fact these tests were also performed with similar results indicating very high agreement between sets of rankings. The result using Kendall's W is reported because of the very specific way in which the null and alternative hypothesis may be conceptualized; a conceptualization with particular relevance for our validation procedure.

H0: There is no association between ratings and a known standard.

H1: Ratings are associated with the known standard.

Using this coefficient in the appropriate statistical test with a null hypothesis of no agreement among the sets of rankings yields p-values of 0.0021 and 0.0022, respectively, and rejection of the null hypothesis. Granted, there could be some objections in formulating the null and alternative hypothesis as above because our ranking ultimately derives from manual scoring of work events, an intermediate step in ordinal ranking. The assumption is that the auditors were using an implicit internal scoring system to arrive at their rankings and that this tacit system is comparable to our scoring system.

As stated, the validation procedure could be made more rigorous by employing a greater number of auditors and a greater number of cases. As an example, the auditors reported that they felt their rankings for two cases (one for each auditor) were not based on their direct experience; i.e., the case notes contained an idiosyncratic situation that they had not encountered before. A wider pool of auditors could help eliminate this situation. The cases should, at a minimum, be re-ranked based on the work boundary used in manual scoring (White 8 CM work) rather than by considering all work for a case by all CMs involved with the case; e.g., work for patients transferring from one floor to another. Changing one ranking based on this update would alter the rankings of other cases, although typically by only +/- 1. At the time of validation, concerns of this type were not an issue because we were more interested in classification of cases by workload rather than assigning a numerical score.

Even better would be to have auditors assign a score to the case after the proposed refinement of the scoring system presented in Table 4.1. The obstacle to more rigorous validation is that reviewing case notes and scoring/ranking cases manually is very time-consuming. Still, even with acknowledged deficits in the scoring system and validation methodology, the manual scoring seems to produce a validated response variable. In the next section this response variable is used to train and validate and automated scoring model/method.

4.5 Developing an automated retrospective scoring methodology

With a validated manual scoring methodology and scores for 1351[6] cases we were in a position to train a model to score cases in a more automated way. As stated, 270 of the 1621 possible cases for the time periods examined had zero notes during their stay on White 8 (or White 9) and are not candidates for automated scoring[7]. If a valid automated scoring methodology can be developed then it would be possible to more easily extend the process of creating a work score response variable based on retrospective analysis of case notes to other floors. In turn, the ability to measure work across floors, patients, and days, as well as various combinations of these possible perspectives, can potentially provide the information necessary to make better informed staffing decisions from a baseline static perspective and a near real-time dynamic perspective[8]

[6]Only 1351 of 1621 cases examined had any documented CM work.

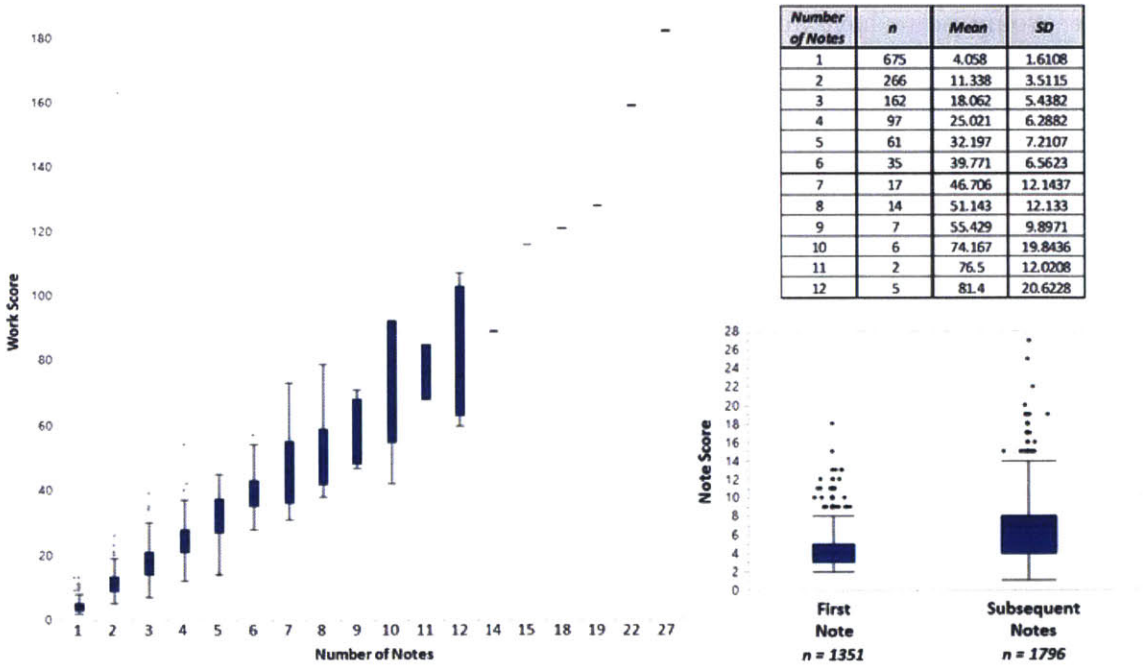
[7]These 270 cases, denoted "zero" workload cases would score as a zero with an automated method.

[8]Of course, the ability to dynamically alter staffing patterns on short timescales requires a model to predict daily workload, the subject of Chapter 6.

This section is divided into five subsections and lays out the rationale and support for the automated scoring methodology in section 4.5.4. With a response variable, the first subsection considers a linear model, using the number of notes for a case as the regressor, to predict work score. As will be shown, this simple model performs well under certain conditions, though ultimately suffers from shortcomings that make it untenable. However, this subsection is of great importance for developing the rationale for a refined model using text-analytical techniques that depend on a text feature vector for modeling. The second subsection describes one of the foundations for the final automated scoring model. Section 4.5.3 outlines the entire sequence of steps used to process case note text in order to form a text feature vector. Some of these steps have utility in other domains, while others are specific to the problem at hand. Section 4.5.4 describes the automated scoring methodology using a regression tree to operate on the feature vector and summarizes automated scoring performance.

4.5.1 Evaluating a simple linear model to automatically score patients

It is reasonable to assume that a strong linear relationship exists between the number of notes for a case and the score for a case. Indeed, such a linear relationship is virtually guaranteed and borders on tautological in nature; since a case is scored based on work events documented in case notes more notes should imply more work events which, in turn, is correlated with a higher work score for a case. Figure 4-12 shows the distribution of work scores for a given number of case notes and clearly indicates a strong linear relationship[9].



Number of Notes	n	Mean	SD
1	675	4.058	1.6108
2	266	11.338	3.5115
3	162	18.062	5.4382
4	97	25.021	6.2882
5	61	32.197	7.2107
6	35	39.771	6.5623
7	17	46.706	12.1437
8	14	51.143	12.133
9	7	55.429	9.8971
10	6	74.167	19.8436
11	2	76.5	12.0208
12	5	81.4	20.6228

Figure 4-12: Distribution of case work scores for different numbers of case note

In fact, Figure 4-12 implies that a simple linear model may give a very accurate score for certain patients. Yet, the side panels in the figure should give one pause when considering the possible

[9]Only 1347 of 1351 cases, those with less than 14 notes, are shown in the left panel of the figure.

performance of a linear model over the range of notes that may be present for a case. Wide-ranging high performance requires that the score value of a note exhibits a relatively narrow range. In fact, even looking at the score distribution for the first note of cases versus subsequent notes suggests that this condition does not hold. Similarly, the generally increasing standard deviation for a case score as the number of notes increases, shown in the table, suggests that this condition is not likely to be met.

Figure 4-13 shows the fit, over the entire data set, for a linear model of case work score with number of notes as the sole regressor. A cursory examination of this figure likewise reveals a strong linear relationship between the number of notes and the work score of a case. Over the entire data set the R^2 for the model exceeds 0.92 with a RMSE of 4.53. Thus, both relative and absolute measures indicate a high model performance. The 95% confidence and prediction intervals are very narrow. What is more, performance metrics remain remarkably identical for cross-validated models or models tested and validated using the holdout method for out-of-sample tests.

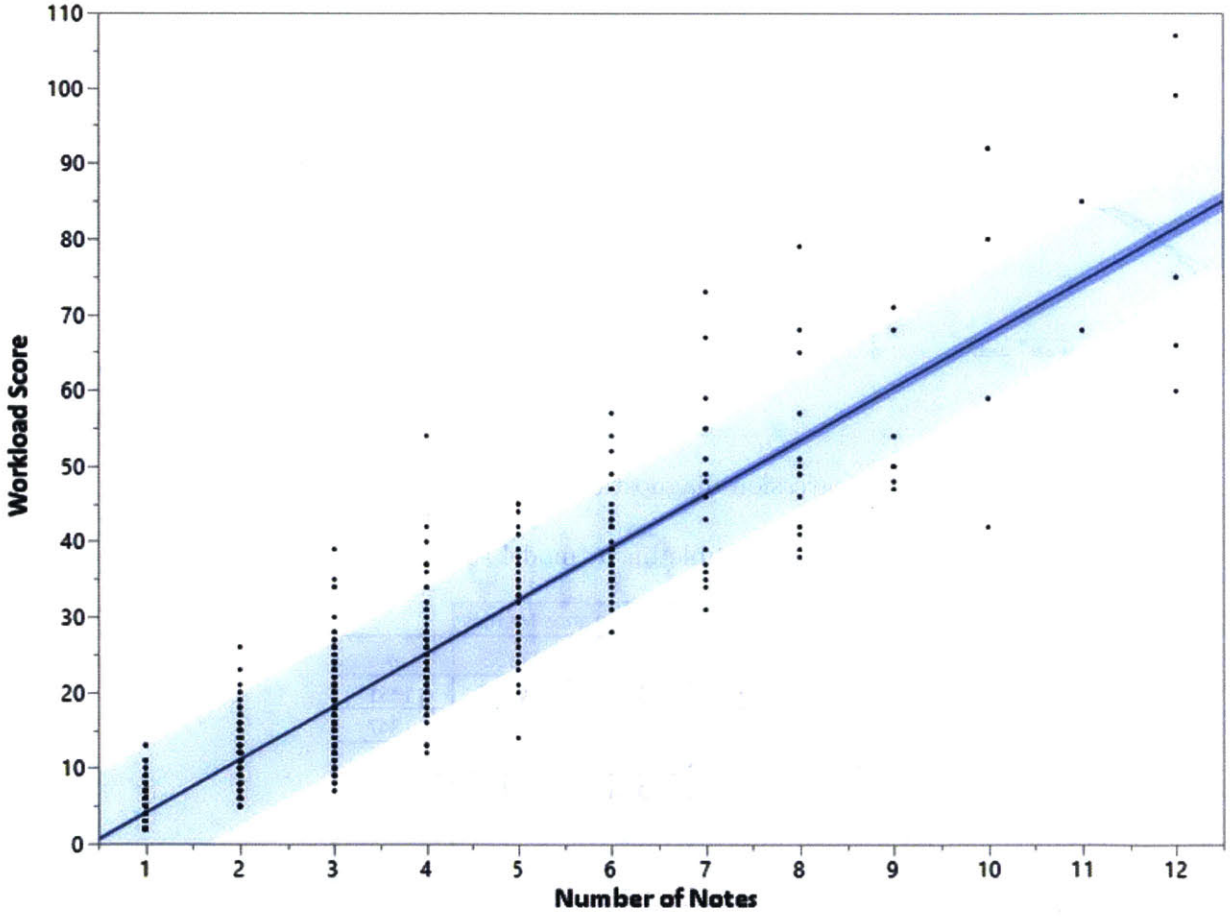


Figure 4-13: Fitting a simple linear model, n = 1347, R^2 = 0.924, RMSE = 4.53

However, a thorough analysis of the model reveals that the high observed performance is due to the large number of cases with only a single case note. Just as including the 270 cases with zero notes and a work score of zero would inflate the apparent performance of a linear model with number of notes as the regressor, so too does the presence of so many single note cases. That is, the model

predicts case score very well for a low number of notes. Performing regression diagnostics illustrates the shortcomings of the model more vividly. Figure 4-14 provides some diagnostic plots, specifically looking for non-normal errors and heteroscedasticity.

The first panel of the figure shows the normal quantile plot for the residuals and exhibits marked kurtosis. The second panel, a plot of the residuals versus the predicted work score for a case shows the fan pattern indicative of heteroscedasticity. The third panel illustrates this heteroscedasticty even more clearly. Here the predicted scores were split at the median and the corresponding residuals were examined. The plot clearly shows heteroscedasticity and, likewise, any statistical test for unequal variances (e.g. O'Brien, Brown-Forsythe, Levene, Bartlett, etc.) indicates an unequal variance of error across the prediction range. Table 4-2 illustrates this behavior in another way by restricting the model to a subset of cases with an increasing note minimum. The performance metrics of a linear model plummet as the minimum number of case notes considered is increased. What is more, this table is for the entire sample of cases at higher minimum note levels; splitting the data set to perform out-of-sample testing results in even poorer performance.

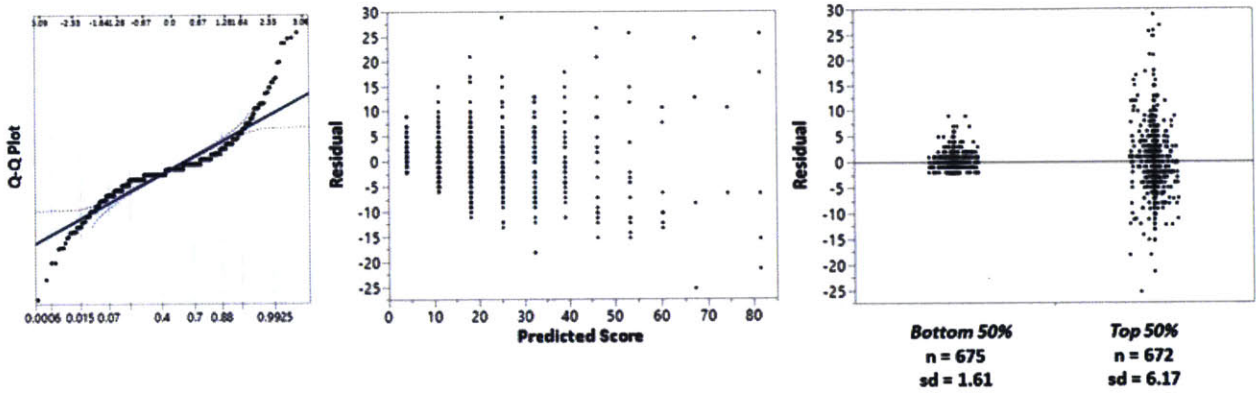


Figure 4-14: Regression diagnostics for the simple linear model

Table 4.2: Decreasing performance of a simple linear model as total note count increases for a case

Note Range		Performance Metrics		
Min	*Max*	R^2	*RMSE*	*n*
1	27	0.92	4.53	1351
1	12	0.89	4.51	1347
2	12	0.82	6.17	672
3	12	0.76	7.42	406
4	12	0.70	8.52	244
5	12	0.64	9.75	147
6	12	0.55	11.27	86

4.5.2 Developing the foundation for an automated scoring procedure with refined differentiation of note types

Clearly, a simple linear model is lacking and can be discarded, but considering why a linear model fails suggests techniques that, if feasible, allow the necessary model refinement. Also, the problem

of heteroscedasticity is not likely to be confined solely to a linear model and must be examined regardless of the predictors and form of the final model.

The fundamental reason why the linear model fails is that a single note may take on a wide range of work scores depending on the number and type of events documented in the note. A simple linear model also suffers because of the limited number of discrete values the number of case notes can assume. This realization, as obvious as it seems, prompted a change in methodology from scoring the complete text from the aggregated notes for a case to scoring each note individually and summing the individual note values to produce an overall case score. An augmented bag-of-words (BOW) model is used to help construct the text feature vectors used in automated scoring.

The limitations of a BOW model were briefly discussed in Chapter 3, but the primary limitation is not the loss of relationships among words; rather it is the loss of context that results. Consider, in some applications analysis using a BOW model performs very well. Two common examples in text-analytics primers are twitter sentiment analysis and analysis of call center transcripts using a BOW approach (see Chapter 3). The typical approach in sentiment analysis is mining tweets about a topic to look for "positive" and "negative" words. The context is clearly established, either by searching tweet text for the existence of keywords or, even better, by a hashtag. Similarly, transcripts from a call center engaged in customer service support for a range of products contains an embedded context – the product that is the subject of the call – that allows matching the number and extent of calls to specific products or incidence of specific problems with specific products.

However, the context for case notes using a BOW approach to score cases is not granular enough when the entirety of text for a case is scored simultaneously. At the note level it is possible to provide a greater level of context to facilitate automated scoring. What is more, the appropriate context can help alleviate some of the problems associated with stylometry effects between different note authors. Scoring at the note level also makes it possible, as in Chapter 5, to divide the aggregate work for a case across the days when the work was actually performed. Granted, there are possible methodological objections to consider because, technically, validation of the manual scoring methodology occurred at the patient/case-level, not the note level. However, the performance gains that can be demonstrated by scoring at the note level, and summing notes to get a case score, or even a daily score (sum of work for all cases in a given day) far outweigh any objections.

At the note-level the primary context can be provided by the type of note; 28 note types were identified in the corpus consisting of the 3147 notes examined. These note types are listed in Table 4.3. Table 4.3 provides various features for each identified note type, each of which is described below. The significance of these columns in relation to automatically assigning a note type header follows the summary descriptions.

The first column provides a description for the note type, while the second provides the header designation used in the text feature vector described next. The third column provides the higher-level group, if it exists, for the note. Some note types have textual markers in common so multiple markers must be examined to distinguish between, for example, a v2 versus a v2ra note, both of which are members of the v2group. As indicated in column four, some notes have a native header (or implied native header). For example, ift-type notes begin with the phrase "INPATIENT FACILITY TRANSFER". An example of an implied header would be the high risk initial assessment type 2. The note does not contain any text explicitly stating that the note is a high risk initial assessment

Table 4.3: Identification and characteristics of different note types

Note Type / Description	Header	Group	Native Header	Example Textual Marker(s)	Template Text	Possible Work Events	Typical Score Range
High risk initial assessment type 2	v2	v2group	Y	"Patient Name:"	Y	chart review, meeting with patient	5-7
High risk initial assessment type 2 from recent admission	v2ra	v2group	Y	"Patient Name:", "Information taken from a recent admission..."	Y	chart review, meeting with patient	4-6
High risk initial assessment, alternate form	iaother		Y	"CASE MANAGEMENT INITIAL ASSESSMENT NOTE"	Y	chart review, meeting with patient	4-7
High risk assessment with no need for CM intervention	afa	afagroup	Y	";however, after further assessment..."	Y	chart review	3-4
High risk assessment with no need for ongoing CM intervention, initial work required	afaplus	afagroup	Y	";however, after further assessment...", *additional text following* "patient does not require case management intervention at this time"	Y	chart review, requesting order for social work	3-5
High risk assessment, patient does not meet criteria for high risk	dnmc	dnmgroup	Y	"patient does not meet criteria for further case management intervention..."	Y	chart review	2-3
High risk assessment, patient does not meet criteria for high risk, initial work required	dnmcplus	dnmgroup	Y	"patient does not meet criteria for further case management intervention...", *additional text following* "intervention at this time."	Y	chart review, email to Patient Financial Services	3-4
High risk assessment completed with initial assesment pending future completion	wbc		Y	"initial assessment will be completed."	Y	chart review	2-3
Screening for post-MGH inpatient facility placement	ifscreen	ifscreengroup	Y	"INPATIENT FACILITY SCREENING"	Y	meeting with patient, placing referrals	6-7
Screening for post-MGH inpatient facility placement, additional work event required	ifscreen2	ifscreengroup	Y	"INPATIENT FACILITY SCREENING", "spoke with"	Y	meeting with patient, placing referrals, speaking to facility	7-8
Screening for post-MGH inpatient facility placement , multiple additional work events required	ifscreen3	ifscreengroup	Y	"INPATIENT FACILITY SCREENING", "spoke with", "requested I fax"	Y	meeting with patient, placing referrals, speaking to facility	8-12
Transfer to post-MGH inpatient facility	ift	iftgroup	Y	"INPATIENT FACILITY TRANSFER"	Y	meeting with patient, placing referrals, speaking to facility, faxing to facility	6-8
Transfer to post-MGH inpatient facility, additional work event required	iftplus	iftgroup	Y	"INPATIENT FACILITY TRANSFER", "informed the patient's daughter"	Y	brief meeting with patient, arranging ambulance transportation, basic coordination with care team	7-10
Transfer to post-MGH inpatient facility, multiple additional work events required	iftplus2	iftgroup	Y	"INPATIENT FACILITY TRANSFER", "informed the patient's daughter", "spoke with admissions at"	Y	brief meeting with patient, arranging ambulance transportation, basic coordination with care team, phone call to patient's family	8-12
Coordination with psychiatric CM for inpatient psychiatric placement	psychcm		Y	"psych CM", "Psych consult case management note", "Initial Psychiatric Consultation..."	Y	emailing psychiatric CM, discussing case with psychiatric CM	3-5
Referral for visiting nurse association or other home services	vnaref		Y	"VNA REFERRAL PLACED"	Y	meeting with patient, placing referral, contacting VNA liaisons	5-7
Weekly update for a patient not ready for discharge	wup		Y	"WEEKLY UPDATE"	Y	chart review	3-5
Patient accepted to a facility	acceptnote		N	"has been accepted"	N	meeting with patient	4-7
Patient denied by a facility	denynote		N	"denied", "denial", "denials", "declined"	N	meeting with patient, referrals, contacting facilities	5-9
Patient is ready to discharge home today	homenote	treportgroup	N	"is discharging home today", "is ready for discharge home"	N	phone calls / coordination with post-MGH care agencies, ordering DME	5-9
Meeting with patient and/or other members of care team	metnote		N	"met", "meeting"	N	meeting with patient	5-10
Patient's scheduled discharge has been cancelled	nreadynote		N	"not ready for discharge", "discharge is on hold"	N	cancelling transportation, calling facility	3-7
Patient has been referred to a post-MGH inpatient facility or home service provider	refernote		N	"Patient has been referred to..."	N	placing referrals	4-7
CM spoke with patient, patient's family, or members of extra-MGH care team	spokenote		N	"spoke"	N	meeting with patient, extensive phone call with a patient's family member	4-9
Team reports that patient needs a certain level of post-MGH care or previously discussed discharge plan can be executed	treport	treportgroup	N	"Team reports the patient needs rehab...", "Per team, patient is ready for rehab."	N	meeting with patient, providing a list of facilities to patient, placing referrals	5-8
Patient chart reviewed following transfer from another floor or case discussed with care team member	reviewnote		N	"Patient transferred", "case discussed with", "chart reviewed"	N	chart review, phone call, meeting with care team member	4-8
Free-text note exceeding 30 words with dictionary words present	ftxt	ftxtgroup	N	*variable*	N	*variable*	6-12
Free-text note less than 30 words or with no dictionary words present	pnote	ftxtgroup	N	*variable*	N	*variable*	2-4

type 2 but each of these v2-type notes has a textual marker, "Patient Name: "[10].

Some notes have neither an explicit or implied native header. For these notes, note type designations were created. The type for these notes is determined by parsing the first sentence of the note and searching for the identifying textual markers. For notes with textual markers in the first sentence indicating multiple possible note types a priority system[11] is used to designate a single note type header. Examples of textual markers are shown in the table.

Some note types, as indicated in column 5 of the table have characteristic "filler" or templated text. This is text that has words that typically indicate a CM work event but, in the case of filler text, do not. As an example, ift-type notes have the phrase:

> At time of discharge: MGH PHYSICIANS 1) Discharge paperwork and the physician's discharge Summary - must be complete in CAS/POE (this is required paperwork - patient CANNOT be transferred to a facility without this paperwork completed and patient must have a copy sent to rehab at time of transport). 2) Comfort Care Forms - required for all DNR/DNI patients traveling via chair car or ambulance. 3) Narcotics Prescriptions - required on all patients being transferred to SNF who will treated with narcotics. (Patient MUST travel to the SNF facility with a hard copy of narcotics prescriptions) MGH STAFF NURSES 1) A copy of the discharge referrals and physician's discharge summary must ALWAYS be sent with the patient to the facility. 2) If rescheduling the transfer time, please notify the MGH Case Manager of the change and contact the ambulance company x36886 (if traveling via ambulance or chair car) 3) Comfort Care Forms and Narcotics Prescriptions (see above) - attach to discharge paperwork if applicable. Case management continues to be available for consultation and reassessment, if indicated.

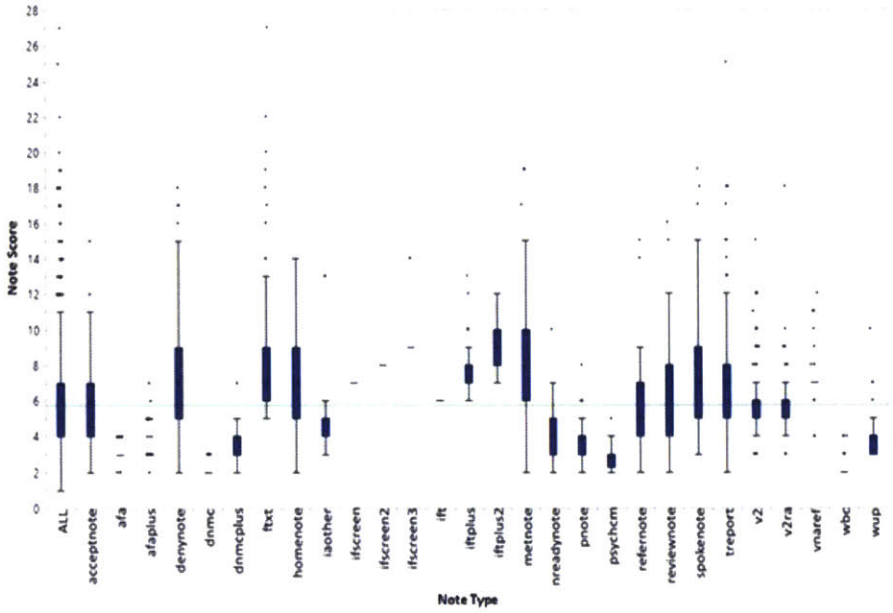
In a BOW model many of the words in this filler text would be indicative of a work event and lead to erroneous scoring of the note. Hence, this text is removed before employing the augmented BOW approach described. Finally, the table lists common work events and score ranges for the different note types.

As an illustration of the benefit of conducting scoring at the note-level and identifying note types, consider Figure 4-15. This figure provides the score distributions for the different note types. The "ALL" distribution is for all notes regardless of type; i.e. without distinguishing note type. Across all non-typed notes the standard deviation is 2.78. By assigning a note type the effective weighted standard deviation across all notes is reduced to 1.99. For the note types highlighted in yellow, comprising 18% of the total notes, the group standard deviation exceeds that for all notes when not distinguished by type. This suggests that further improvements could be made by using additional note subtypes for these high-variance notes.

Though context is the primary benefit of distinguishing note types, typing notes also helps in dealing with stylometry effects between case managers and is indicative of fundamental differences in work

[10]The HRIA2, like other note types, may have multiple characteristic textual markers. In these cases the most accessible textual marker(s) were chosen, such as those with relatively invariant positions within the text.

[11]This priority system could also be refined. Currently a note with multiple possible types is assigned to the highest scoring type based on the mean of notes that can be identified unambiguously. The greatest refinement would likely come from a wider array of note types to lessen the chance of ambiguity in identification.



Type	n	Mean	SD
ALL	3147	5.77344	2.78388
acceptnote	68	5.70588	2.31838
afa	209	3.03828	0.25658
afaplus	184	4.03804	0.61286
denynote	108	7.5	3.51897
dnmc	98	2.17347	0.3806
dnmcplus	32	3.5	0.98374
ftxt	182	7.97802	3.26027
homenote	89	6.98876	2.57388
iaother	108	4.96296	1.59395
ifscreen	39	7.10256	1.04617
ifscreen2	28	8.14286	0.97046
ifscreen3	43	9.2093	1.12458
ift	71	6.4507	1.35001
iftplus	135	7.74815	1.13768
iftplus2	37	8.97297	1.23573
metnote	149	8.26846	2.87024
nreadynote	41	4.26829	1.68856
pnote	188	3.35106	1.01547
psychcm	60	2.88333	0.64022
refernote	28	5.85714	3.07576
reviewnote	79	6.22785	2.86871
spokenote	135	7.46667	3.06351
treport	342	6.97661	2.96005
v2	394	5.52538	1.47778
v2ra	132	5.65152	1.67143
vnaref	51	7.31373	1.37826
wbc	50	2.22	0.58169
wup	67	3.79104	1.14861

Effective standard deviation of all notes when distinguished by note type *decreases from 2.78 to 1.99, a 29% reduction*

Figure 4-15: Distribution of note scores by note type

patterns. Figure 4-16 and 4-17 illustrate these points. This figure shows the note type distribution between the two floors. Put simply, words potentially indicative of work events and event duration are, in fact, more or less indicative depending on the note type. As the figure shows, the prevalence of different notes differs between floors. This accounts, in part, for the different distribution of note scores, also shown in the figure. Distinguishing note types enhances the potential extensibility of automated scoring to other floors and patient populations. On the second point, we can consider common note types between the two floors and their frequency relative to the different sample sizes. For example, dnmc-type notes and wbc-type notes (refer to Table 4.3) are much more common on White 9. As shown in Chapter 5 this increased prevalence is directly related to the differential timing and method of admission window work between the two floors.

4.5.3 Outline of text processing procedures and text feature vector construction and composition

In addition to the note type, the complete text feature vector derived from case notes and used as the predictor for automated scoring has 28 other fields. These fields are briefly described in Table 4.4. Relative to the number of some note types available, the dimension of the feature vector is high. In fact, the automated scoring employing a pruned-regression tree only used three of these fields in a primary manner (Section 4.5.4). Many of the other fields were retained as surrogate nodes but, given that the feature vector for each note was not missing any entries, these surrogate nodes did not come into play. However, all fields are described because in a data set synthesized from actual data that was six times as large (approximately 19000 notes for 8000 patients) many more fields were used for splitting the tree. Likewise, when examining the use of boosted and random forest

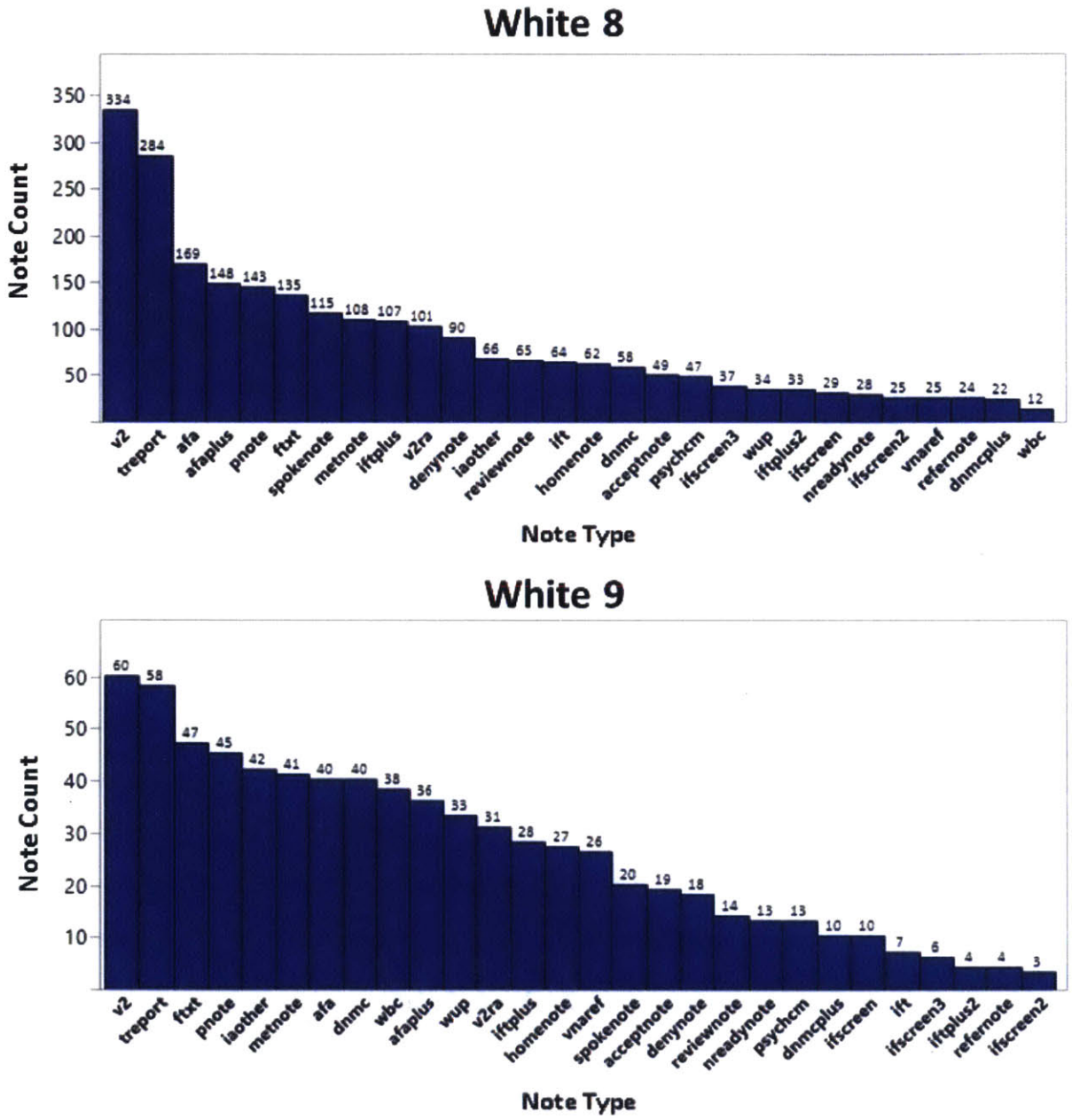


Figure 4-16: Prevalence of different note types on White 8 and White 9

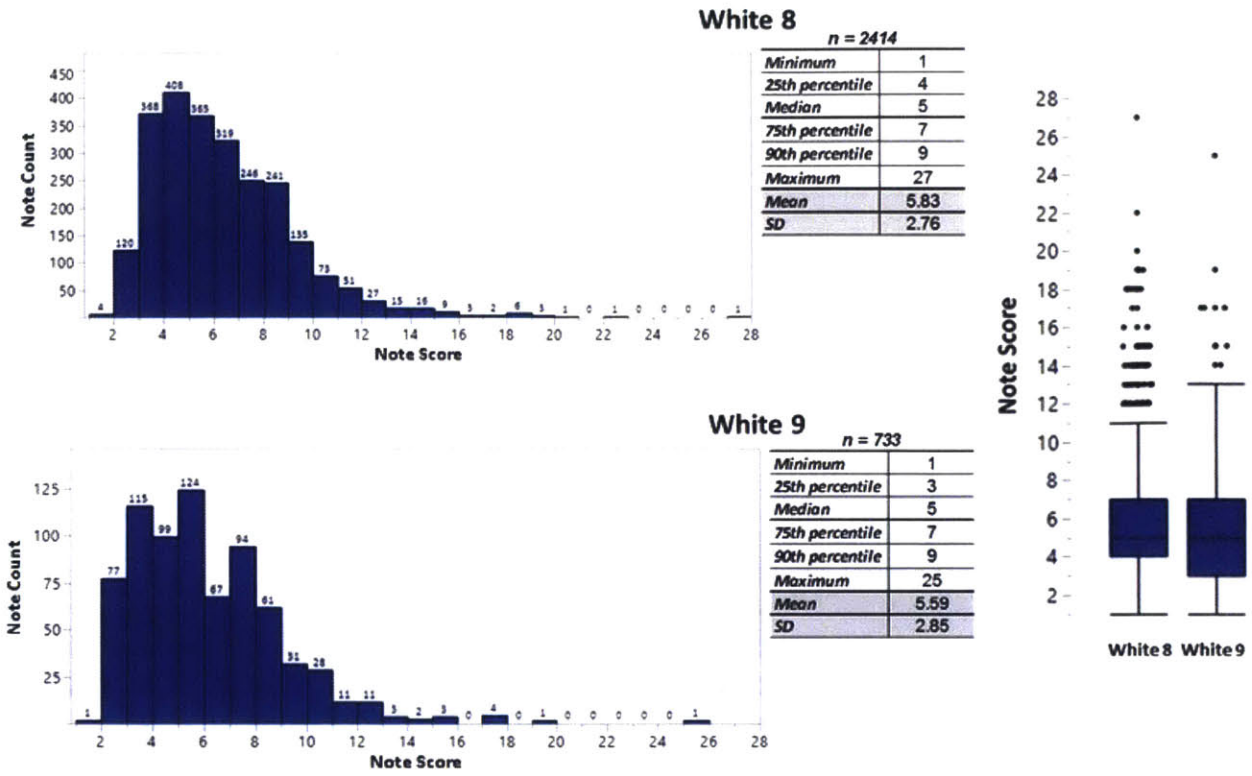


Figure 4-17: Distribution of note scores on White 8 and White 9

regression trees more fields were utilized. However, in the interest of transparency and scrutability a single regression tree was ultimately used. The details of synthesizing patient text data are beyond the scope of this thesis but synthesis required a large amount of data-munging to maintain internal consistency for synthesized case notes. The relevant takeaway is that several fields unused when training and testing with the actual data set assumed prominence with a larger simulated data set. This is also likely to be the case with a larger actual data set. At any rate, the entire feature vector was made available to the model and, given the nonparametric nature of a regression tree and *de facto* inherent feature selection, this did not present issues with modeling.

The three primary fields used were header (note type), wordcount, and previous note. The note-type header was described above. The wordcount was calculated by summing all occurrences of words belonging to the categories listed in the table. Appendix C provides a listing of the words constituting each category.

The previous note type was used by the regression tree to distinguish higher scoring notes. For a patient with a given discharge disposition note types tend to follow a common sequence. When this sequence is broken it may be an indicator of a difficult case and the current note documenting a higher than ususal amount of work given the current note type. Some common sequences are shown in Table 4.5.

As the discussion in this section indicates, some of the techniques to prepare and process the note text were intricate and the devil was very much in the details. Figure 4-18 outlines, at an intermediate perspective, the main phases and steps used to prepare the text and construct a text

Table 4.4: Field descriptions for complete text feature vector

1	header	type of note, primary field
2	met	count of met occurrences in non-metnote or count of occurrences after first sentence of met note; highly indicative of work
3	spoke	count of spoke occurrences in non-spokenote or count of occurrences after first sentence of spokenote; highly indicative of work as primary indicator in contrast to cdwords which relate to duration of work
4	refer	count of words indicating referral placed for patient
5	sword	count of secondary words that may be highly indicative of work but with the caveat that they often document referred actions
6	oword	count of other words that textual analysis of a large, simulated data set and hierarchical clustering suggested no membership to other categories
7	cdword	count of words indication communication or duration of that communication; used to help distinguish where in the scoring range a work event falls (e.g., a limited or extensive meeting or phone call)
8	pword	count of words indicating a patient's preferences; e.g. a home plan or inpatient placement, or which inpatient facility is desired
9	conword	count of words indicating patient's acceptance of a proposed course of action
10	tword	count of trigger words, or words which typically trigger a sequence of actions by a case manager; examples include deny or except; in denynotes and accept notes, for example, this count is for after the first sentence
11	hword	count of words typically associated with a patient discharging home with services
12	rword	count of words that may indicate a work event and are repeated often; this category of words likely varies between case managers
13	senword1	count of words indicating a non-cooperative patient or family; presence tends to accompany longer duration work events
14	senword2	count of words indicating a patient or family that needs a lot of "handholding"/attention by the case manager; also accompany longer duration work events
15	senword3	count of words similar to senword1 count but even more indicative of non-cooperation and long duration work events; like rwords this category likely varies between case managers
16	trword	count of words associated with a patients transportation needs and meeting those needs; these words are also often repeated and frequent
17	wordcount	the second primary field that relates to the relevant length of a note as it is a sum of fields 2-15
18	phase	the phase of a patient's stay during which the note was written: 1 -admission, 2 - pre-discharge planning, 3 - discharge planning, 4 - combined admission and discharge planning
19	dsa	number of days since admission when the note was written
20	dtd	number of days until discharge when the note was written
21	remaining	number of notes remaining for a patient
22	previous	number of notes already encountered for a patient
23	day	day of week when the note was written; notes on Friday and Monday, as well as some notes on the weekend tend to score higher for some notes
24	pphase	phase in which previous note was written for a patient
25	nphase	phase of next note for a patient
26	pnote	previous note type for a patient
27	nnote	next note type for a patient
28	dd	discharge disposition for a patient
29	author	author of note; can be of utility in controlling for stylometry effects across case managers

Table 4.5: Common note type sequences

Patient Description	Typical Note Sequence	Typical Score
Patient admitted from an inpatient facility on a bed hold and in agreement to return	v2 *or* v2ra, ift	10
Patient admitted from home, requiring a post-MGH inpatient facility, no prior subacute facility admissions, sole decision maker	v2, ifscreen2, ift	20
Patient admitted from home, requiring a post-MGH inpatient facility, no prior subacute facility admissions, multiple decision makers	v2, ifscreen3, iftplus	23
Patient admiited from home with VNA services already active	v2, homenote	8
Patient admiited from home, requiring post-MGH home services with no services in place	v2, vnaref, pnote	15
Patient admiited from home, requiring post-MGH home services from multiple providers with no services in place	v2, vnaref, ftxt	20
Patient determined to require inpatient psychiatric admission upon initial admit	v2 *or* afaplus, psychcm, psychcm	11

Figure 4-18: Outline of steps used for text preparation and text feature vector construction

feature vector for each note[12].

Phase II, III, and IV have already been discussed at the appropriate level of detail, except steps III.2 and III.4. Considering III.2, in reading the notes there are several common abbreviations that were replaced to make requisite text recognition easier. For example, ftm is used to refer to family-team meeting, tc for telephone call, vm for voicemail, etc. The list of abbreviations encountered is certainly not exhaustive so this step would not be comprehensive for novel text. Step III.4 is important to prevent assigning a higher score than warranted at the note-level. Referred actions were discussed previously – when a word typically indicative of a CM work event actually refers to work completed by another individual. With a BOW model these referred actions can be difficult to distinguish. In the case of referred actions that are part of filler text in notes with native headers, explicit or implied, these referred action words are eliminated when the filler text is stripped. For non-native header notes removal is more hit-or-miss. There are some textual markers that allow referred action words to be discarded. The most common of these is a sentence beginning with the word "Please" indicating that the case manager is requesting action from another member of the care team.

There is still work that needs to be done to fully automate retrospective scoring of case notes, primarily involving coding to more effectively process text and remove desired text with the use of regular expressions. Qualitatively, on a scale of 1 to 5, with 5 being fully automated, the overall automation level is 3. A qualitative assessment for all phases and steps is provided in Table 4.6.

Some of the steps in Phase I are particularly difficult to automate. In Phase I, with respect to patient movement between floors there are four types of patients. Considering the specific case of White 8, these four patient types are:

1. Patients with White 8 as the first inpatient unit and discharge unit
2. Patients with White 8 as the first inpatient unit and a different discharge unit
3. Patients with White 8 as the discharge unit and a different first inpatient unit
4. Patients that transit White 8, having a different first inpatient unit and discharge unit

The available data allows determination of when a patient enters and leaves a floor, assisting in the allocation of work score to the correct floor. However, the CM for the losing floor may batch documentation tasks and submit a case note after the patient has already transferred to the gaining floor. This can be a difficult condition to identify. This delayed documentation can also occur after a patient discharges. Furthermore, a CM may either record events for multiple days in a single note or may record events that occurred yesterday on the current day. In some cases there are embedded time stamps as part of the text written by the CM, rather than system time stamps, that allow work to be allocated to the correct day. In other cases there are textual markers such as "ftm held yesterday" that aid in this process. Finally, in some cases a note is completed on one day and then edited at a later date. These notes could have as many as three time stamps making it problematic to automate the inference necessary to allocate work to the correct day and not double count work.

[12]Some of the steps shown are legacy steps from earlier versions of our text-analytical engine. For example, there is little need to remove numbers or special characters in the current version, except for the fact that it makes manual

Table 4.6: Assessment of automation-level for case scoring

Phase I	*Allocate note to correct patient, day, and floor*	Overall automation = 2
Step(s)	Automation Level	Challenges
1	5	
2	4	Delayed documentation documentation on one floor results in time stamp leading to erroneous assignment of work to another floor for patients with inter-floor transfers
3	3	Delayed documentation documentation on one floor results in time stamp leading to erroneous assignment of work to another floor for patients with inter-floor transfers
4	5	
5	2	Events from multiple days doumented in a single note
6	2	Multiple time stamps with "correct" time stamp embedded in note text or note with today's time stamp documenting previous days events

Phase II	*Determine and append note-type header to each note*	Overall automation = 3
Step(s)	Automation Level	Challenges
1	5	
2	4	Template has minor modifications between authors that present major difficulties when using regular expressions to identify text strings to strip
3	2	BOW model ill-suited to distinguish between direct CM action and referred actions, difficulty determining exact number of additional work events used to distinguish subtypes
4	5	
5	3	Parsing sentences is not 100% accurate
6	5	
7	4	Criteria used for distinguishing may result in errors for unseen text if work event words not in current dictionary are encountered
8	5	

Phase III	*Text-processing and forming note-term matrix*	Overall automation = 4
Step(s)	Automation Level	Challenges
1	5	
2	5	
3	5	
4	2	Inter-CM and intra-CM syntax for referred actions varies greatly; requires higher-level natural language processing and parts-of-speech labeling to fully automate
5	5	

Phase IV	*Construct text feature vector for each note*	Overall automation = 3
Step(s)	Automation Level	Challenges
1	5	
2	5	Template has minor modifications between authors that present major difficulties when using regular expressions to identify text strings to strip
3	5	BOW model ill-suited to distinguish between direct CM action and referred actions, difficulty determining exact number of additional work events used to distinguish subtypes
4	5	
5	2	Field 17, Phase, presents difficulty when distinguishing between
6	5	

4.5.4 Automated scoring via regression tree and model performance

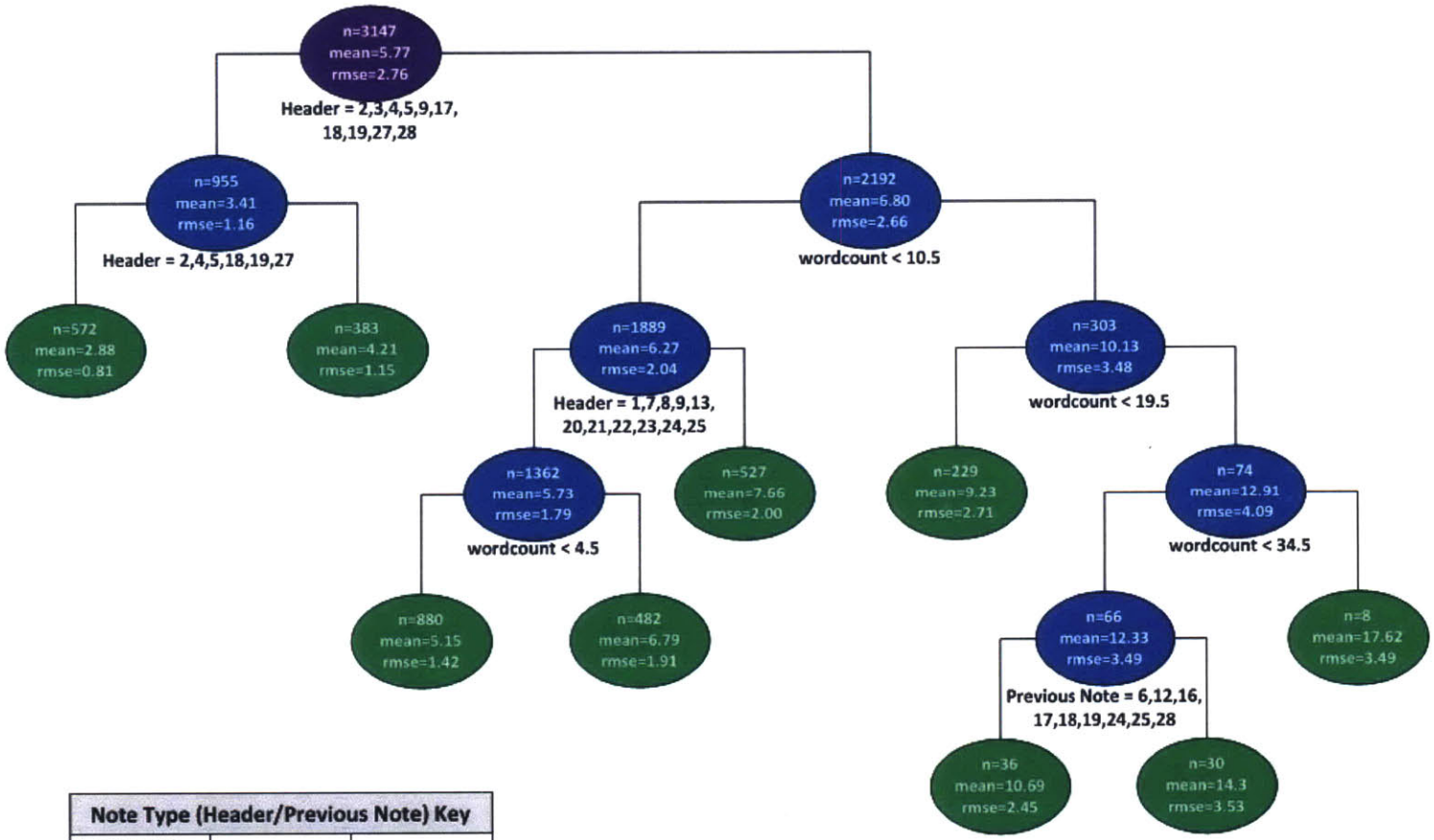
As alluded to, automated scoring revolved around the use of a regression tree, with a standard CART implementation, operating on the text feature vector to score individual case notes[44]. The scores of the individual notes for each case were then summed to get a case score. From the daily perspective, the scores for all notes on a given day were summed to get a daily score.

The number of splits was controlled by tuning the complexity parameter and the minimum bucket size in a grid search. The complexity parameter provides the threshold by which a split must decrease the relative error (or, equivalently, increase R^2) for the split to be attempted[124]. The minimum bucket size is the minimum size of a terminal node that can result from a possible split. Of these two parameters, the complexity parameter tends to have vastly greater importance for splitting nodes.

For a given value of the complexity parameter and minimum bucket size employing a grid search, trees were grown with an increasing number of splits (terminal nodes = number of splits + 1) and the 5-fold cross-validated error. The value of the complexity parameter and minimum bucket size resulting in the lowest cross-validated error was used to select a fully-grown tree with the corresponding number of splits. The tree was then pruned to the minimum number of splits within 1-standard error of the minimum cross-validated error[13]. Figure 4-19 shows a representative tree grown and pruned using the full data set. The tuned, pruned tree used a complexity parameter of 0.8, minimum bucket size of 7 and resulted in 10 splits (9 terminal nodes). As discussed below the performance on the full data set is representative of performance with various and separate training

scanning of encoded notes easier. Similarly, with dictionary encoding of the note text into word categories, replacing known abbreviations is an unnecessary step.

[13]Given the care taken in pre-processing the text this may be an overly conservative pruning procedure.



Note Type (Header/Previous Note) Key		
1-acceptnote	11-ifscreen2	20-refernote
2-afa	12-ifscreen3	21-reviewnote
3-afaplus	13-ift	22-spokenote
4-dnmc	14-iftplus	23-treport
5-dnmcplus	15-iftplus2	24-v2
6-denynote	16-metnote	25-v2ra
7-ftxt	17-nreadynote	26-vnaref
8-homenote	18-pnote	27-wbc
9-iaother	19-psychcm	28-wup
10-ifscreen		

Figure 4-19: Representative form of regression tree for automated scoring

and validation sets.

In growing the tree 13 fields of the text feature vector were used, while in the pruned tree only 3 fields were used – header, wordcount, and previous note. The relative importance of the 13 variables used to grow the tree is shown below in Figure 4-20[14]. As a means of relating variable importance to the effect of primary splits, the first split on header reduced the relative error on the order of 32%, the next split on header reduced relative error by 16%, while the third split, on wordcount, reduced the relative error by 7%. The remaining five splits reduced the relative error a combined 9%. Of course, the exact nature of improvement with subsequent splits depends on the specific training set used. Still, the importance of wordcount, and, especially, the note type header on scoring performance is clear.

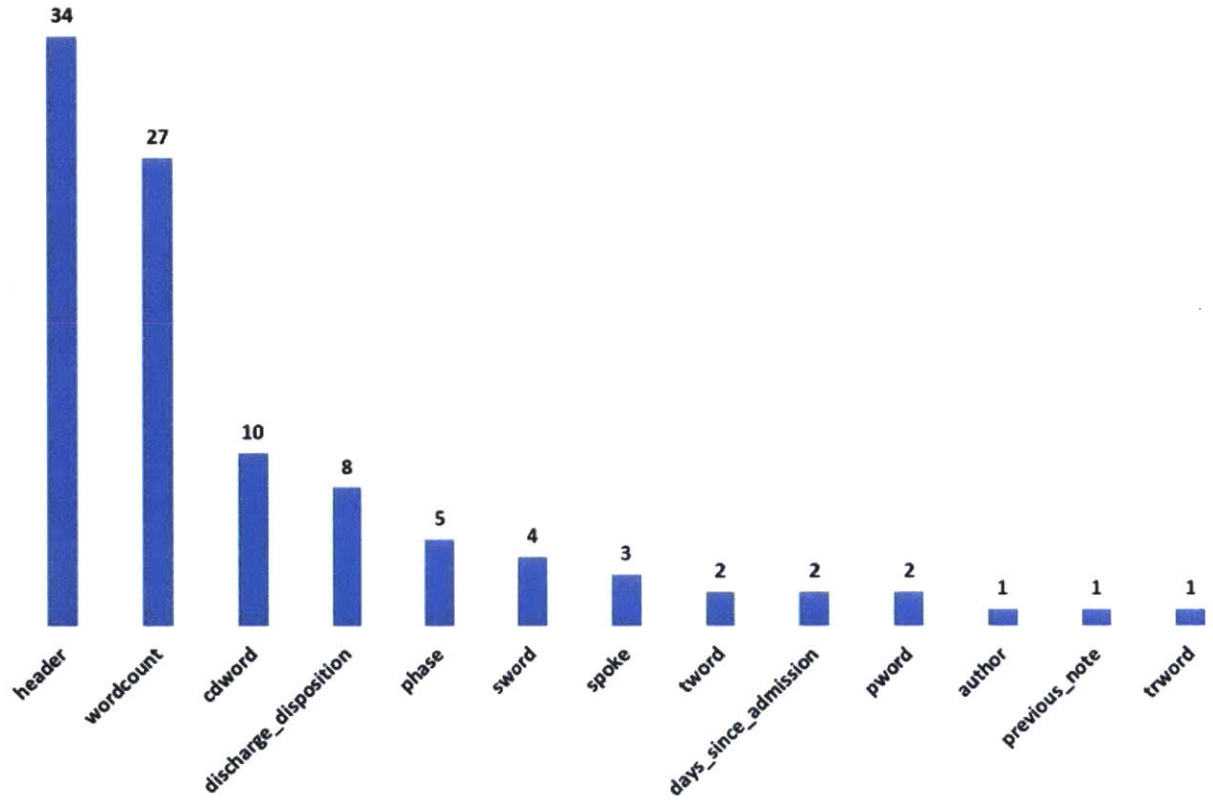


Figure 4-20: Relative feature importance for regression tree

As one can see by looking at a typical tree in Figure 4-21, the effect of pruning was to eliminate those terminal nodes used to classify relatively high scoring notes. Figure 4.5.9 shows a plot of the predicted note score versus actual (manual) note score. At the note level the performance of scoring is has an $R^2 = 0.63$ and RMSE = 1.69. An individual note is limited to 9 values at which it can be scored, the mean response of each terminal node. However, the important scale of performance

[14]This if for illustration purposes only and the y-axis units have absolute meaning, only relative meaning. See Therneau and Atkinson's vignette for further exposition how relative importance is calculated in this figure[124].

is at the patient and day level. By scoring at the note level and summing notes to get scores at higher levels, we can get good model performance[15]. Figure 4-22, comparing the linear model and the regression tree model at the patient level, plotting actual case score by predicted case score, clearly demonstrates this aspect. In this figure the red line is a 45° line of agreement, not a fit line.

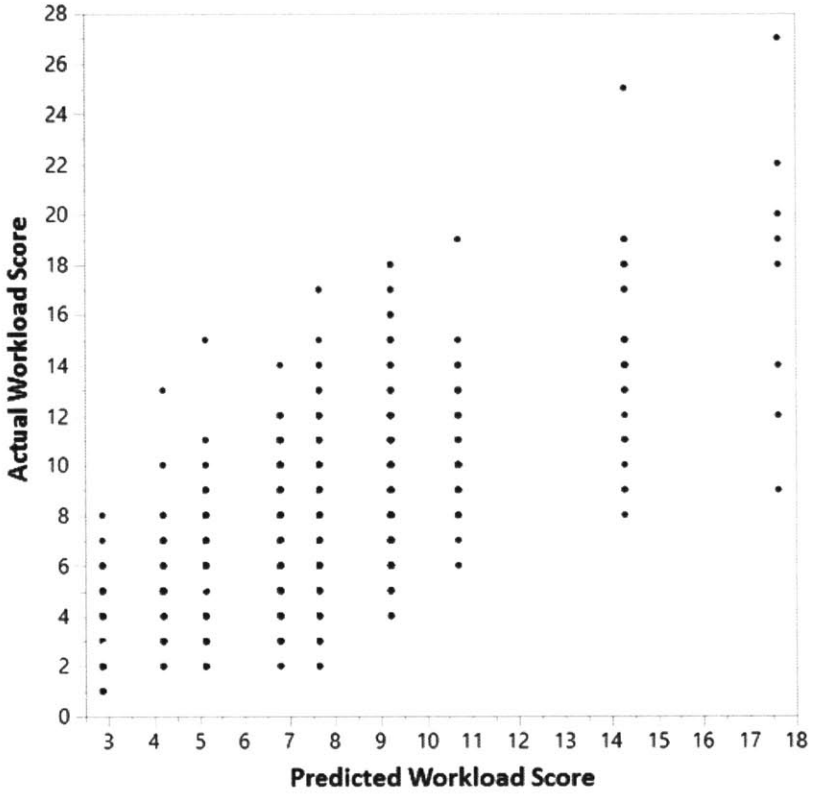


Figure 4-21: Performance of automated regression tree scoring at the note level, full data set used, $R^2 = 0.63$ and RMSE = 1.69

The performance metrics for the regression tree-based model are more properly viewed as imputed metrics since actual modeling occurred at the note level. However, the interpretation in terms of performance is the same. From a practical standpoint the performance metrics indicate that the core of the automated scoring method is valid.

Of course, evident in Figure 4-22 is heteroscedasticity. This is almost certainly a result of the pruning algorithm or, more correctly the pruning algorithm given the relatively small sample of higher scoring notes (relative to feature vector)[115]. Pruning tends to eliminate terminal nodes associated with high value (and high variance) notes in our effort to reduce relative and absolute error on novel data. The highest scoring patients tend to have at least one of these high scoring notes in their record. The heteroscedasticity of the tree is compared with the linear model in Figure 4-23. The extent of heteroscedasticity has clearly been reduced, but the effect is still present.

[15]The better performance, compared to the linear model, at the case level stems from the fact that cases with a single note are more accurately scored when identifying the specific note type. This also applies to cases with an intermediate number of notes. Furthermore, for cases with multiple notes, the different note types tend to occur only once; the errors for the different note types tend to move in different directions rather than being reinforcing. However, heteroscedasticity is still a concern.

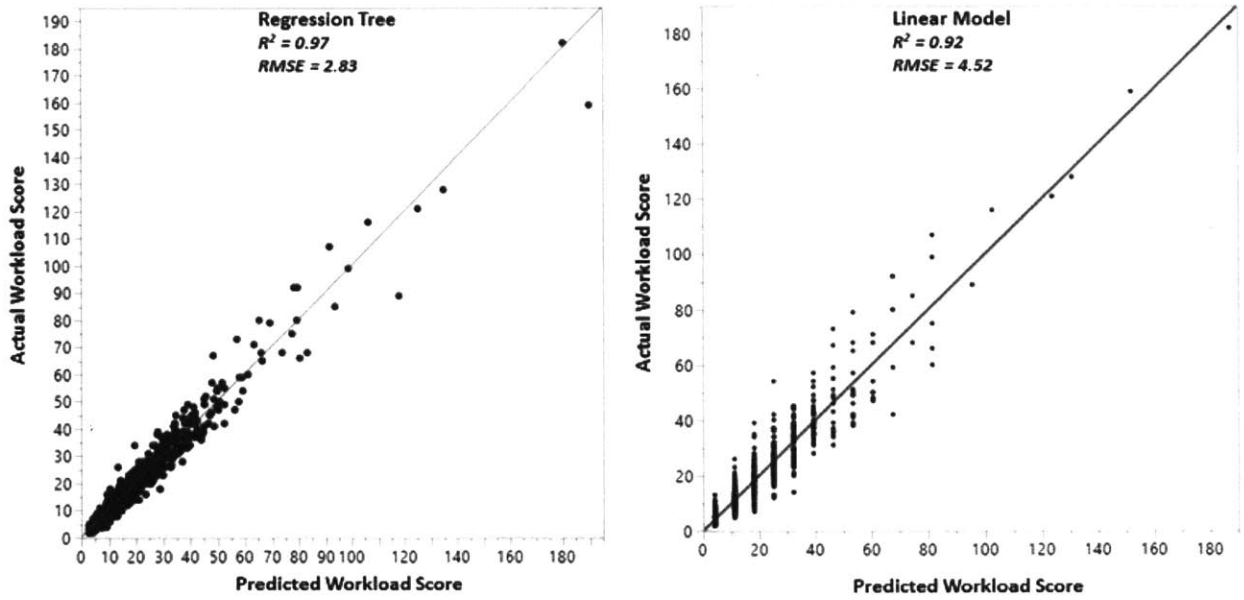


Figure 4-22: Comparison of linear regression and regression tree performance at the patient (case) level

Figure 4-24 shows the performance of obtaining a daily work score by summing the scores for each note associated with that day. Again, the performance metrics are technically imputed values since modeling occurred at the note level but the distinction is of no practical significance. The daily scores and associated variability are considered more fully in succeeding chapters. It bears noting that, from a perception of workload perspective, at some point adding individual note sores to get a daily score would not be valid; linearity would be violated[16]. That is, consider two days each with a work score of 150, with one of the days having 8 active cases and the other having 16 active cases. The day with 16 active cases would undoubtedly be perceived as a higher workload, particularly as CMs switch between all active cases on a given day rather than working each case in succession to daily completion. Currently there is no validated way to account for these type of effects.

All of the preceding discussion of automated scoring performance with a regression tree operating on a constructed text feature vector utilized the full data set. Table 4-2 shows the model performance using separate training and validation sets. For trials with training and validation sets the tree was grown and pruned using the procedure outlined at the beginning of this section using only the training set. The performance of the pruned tree on the test set was then examined. Of course the tree differed from the exemplar in Figure 4-19, but each pruned tree had nine nodes and split on the same features. Also, as shown in the table, the performance at the note, patient, and day level remains remarkably consistent.

For modeling iterations 3-7 ten trials were performed with the average and resulting interval reported for performance metrics. For iterations 3 and 5 the notes were randomly split at the patient level, allowing evaluation of patient level performance; iterations 4 and 6 randomly split notes at the day level. Again, the performance of the model is very consistent. Of note, in iteration 7, the model

[16]This would likely result due to perceived workload effects, independent of objective workload[136][113].

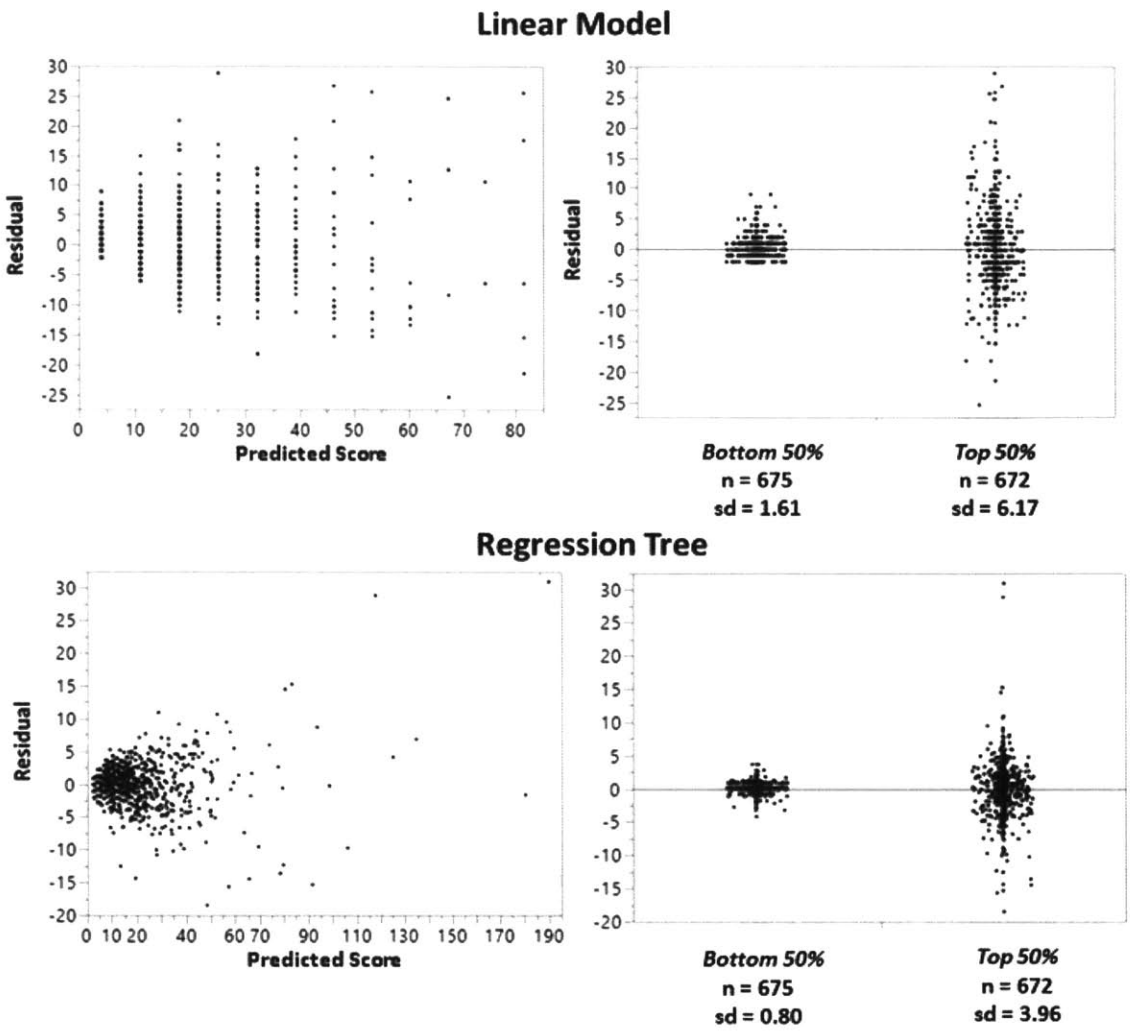


Figure 4-23: Comparing the heteroscedasticity of linear regression and regression tree at the patient (case) level

was trained on White 8 notes and tested on White 9 notes, suggesting that the automated scoring methodology is generally extensible to floors of the same type. Note that the RMSE at the day level for iteration 6 is significantly higher than for other iterations with a training and test set. This is because for the days in the test set during 1 April 2015 – 30 June 2015, the day score consisted of the sum of scores for White 8 and White 9.[17]

To conclude this chapter, there is ample evidence suggesting the automated scoring methodology is valid. Of course, further work needs to be completed for both validation and in fully automating the retrospective scoring methodology. Chapter 5 continues an examination of the current state of work on White 8 and White 9, using the work metric to develop ideas needed to arrive at a preliminary model for prediction of daily workload.

[17]The relative RMSE (relative to the mean response) was comparable.

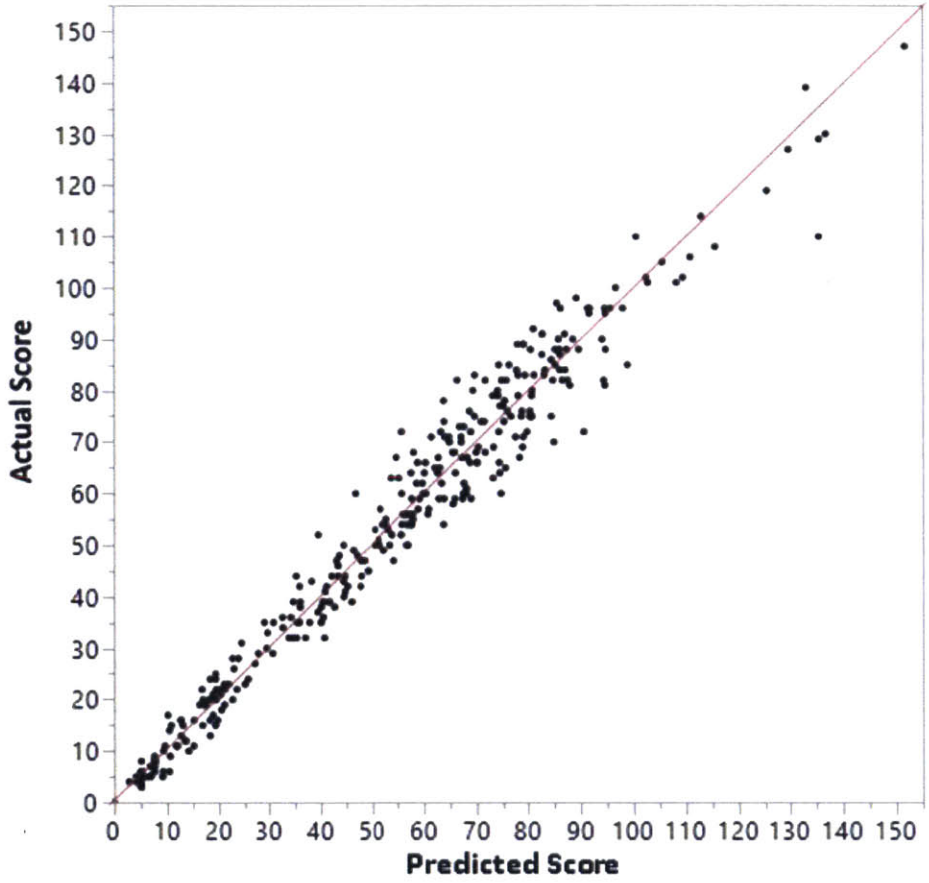


Figure 4-24: Performance of automated regression tree scoring at the day level, full data set used, $R^2 = 0.97$ and RMSE = 5.53

Table 4.7: Performance of automated scoring for various training and validation data sets

				Validation Set Performance					
				Note-level		Patient-level		Day-level	
Model Number	Model Type	Training Set	Validation Set	R^2	RMSE	R^2	RMSE	R^2	RMSE
1	OLS Regression	Fit entire data set				0.89/0.92	4.51/4.52		
2	Regression Tree	Fit entire data set - Pruned tree		0.63	1.69	0.98	2.85	0.97	5.57
3	Regression Tree	75% of patient notes from White 8	25% of patient notes from White 8	0.57 (0.54,0.59)	1.83 (1.77, 1.92)	0.95 (0.94, 0.97)	2.35 (2.30, 2.37)		
4	Regression Tree	75% of daily notes from White 8	25% of daily notes from White 8	0.58 (0.55, 0.59)	1.71 (1.68, 1.80)			0.97 (0.96, 0.98)	5.00 (4.96, 5.09)
5	Regression Tree	75% of all patient notes (White 8 and White 9)	25% of all patient notes (White 8 and White 9)	0.57 (0.55, 0.59)	1.79 (1.76, 1.82)	0.95 (0.91, 0.97)	3.34 (3.28, 3.46)		
6	Regression Tree	75% of all daily notes (White 8 and White 9)	25% of all daily notes (White 8 and White 9)	0.56 (0.53, 0.57)	1.92 (1.89, 2.04)			0.98 (0.96, 0.99)	7.04 (6.87, 7.34)
7	Regression Tree	White 8 notes	White 9 notes	0.60 (0.58, 0.61)	1.81 (1.76, 1.89)	0.97 (0.95, 0.98)	2.62 (2.56, 2.65)	0.97 (0.95, 0.98)	5.20 (5.14, 5.32)

Chapter 5

Current State Analysis of Case Manager Work

Chapter 4 detailed our efforts to develop a meaningful metric for measuring case manager workload and developing a process to automate retrospective scoring. Meaningful, in this sense, means that the metric allows a comparison of the amount of work completed for one patient during their length of stay to the amount of work completed for another patient. Further, by scoring at the note level, it was possible to sum scores to get an indication of the amount of work completed at the day level. Using the metric in this way has utility in itself. For example, with further refinement, such a metric could be the basis for comparing the relative amounts of work done, per patient or aggregated over some time horizon, for different floors or case manager positions. This could, in turn, help to establish rational baseline (static) staffing levels giving the widely varying, by floor, inpatient populations at MGH.

This chapter uses the metric developed and validated in the previous chapter to perform an analysis of the current state of case manager work on White 8 and White 9. Specifically, the distribution of work throughout a patient's length of stay is examined. Understanding this distribution is a prerequisite for a predictive model with even a modicum of utility.

While Chapter 4 is the foundation for our work, Chapter 5 bridges the gap to predicitve modeling. This chapter provides a framework for understanding the various ways a patient may progress, from the case management perspective, from admission to eventual discharge. The underpinnings of this framework are two-fold. On the one hand, various phases through which the patient progresses on the case management / discharge planning plane are more rigorously defined. Then, within this phased sub-framework, common reference modes of work distribution at the patient level are identified and summarized.

The complete framework is then used to examine some important differences in the amounts and, to a lesser extent, distribution of scored work for White 8 and White 9, floors with ostensibly similar patient populations, during the time periods examined. This comparative analysis is important for gaging how easily model(s) developed for one floor can be extended as is or modified for other floors. More fundamentally, explaining any discrepancies between similar floors within the analysis framework is important verification of the framework's utility.

The ultimate aim of this chapter is to provide the rationale for the preliminary predictive model of Chapter 6, as well as the rationale for model improvement recommendations and, more importantly, operational recommendations that may be supported by our work.

5.1 Introductory discussion of case manager work during a patient's LOS

Perhaps the most natural way to view the distribution of work across a patient's LOS, aggregated across all patients, is with respect to patient admission to a floor and subsequent discharge from that floor. Figure 5-1 shows this distribution, both at the aggregate level and average level[1]. As in Chapter 4 , the time period considered for White 8 is 1 October 2014 – 30 June 2015 (1229 patients), while that for White 9 is 1 April 2015 – 30 June 2015 (392 patients). The aggregate amounts of work for each day relative to admit or discharge provide the sum for the patients with work scored on that day; the average is calculated similarly across the patients with work scored on that day (that is, the average is conditioned on work scored for a patient and not across, for example, all 1229 patients on White 8).

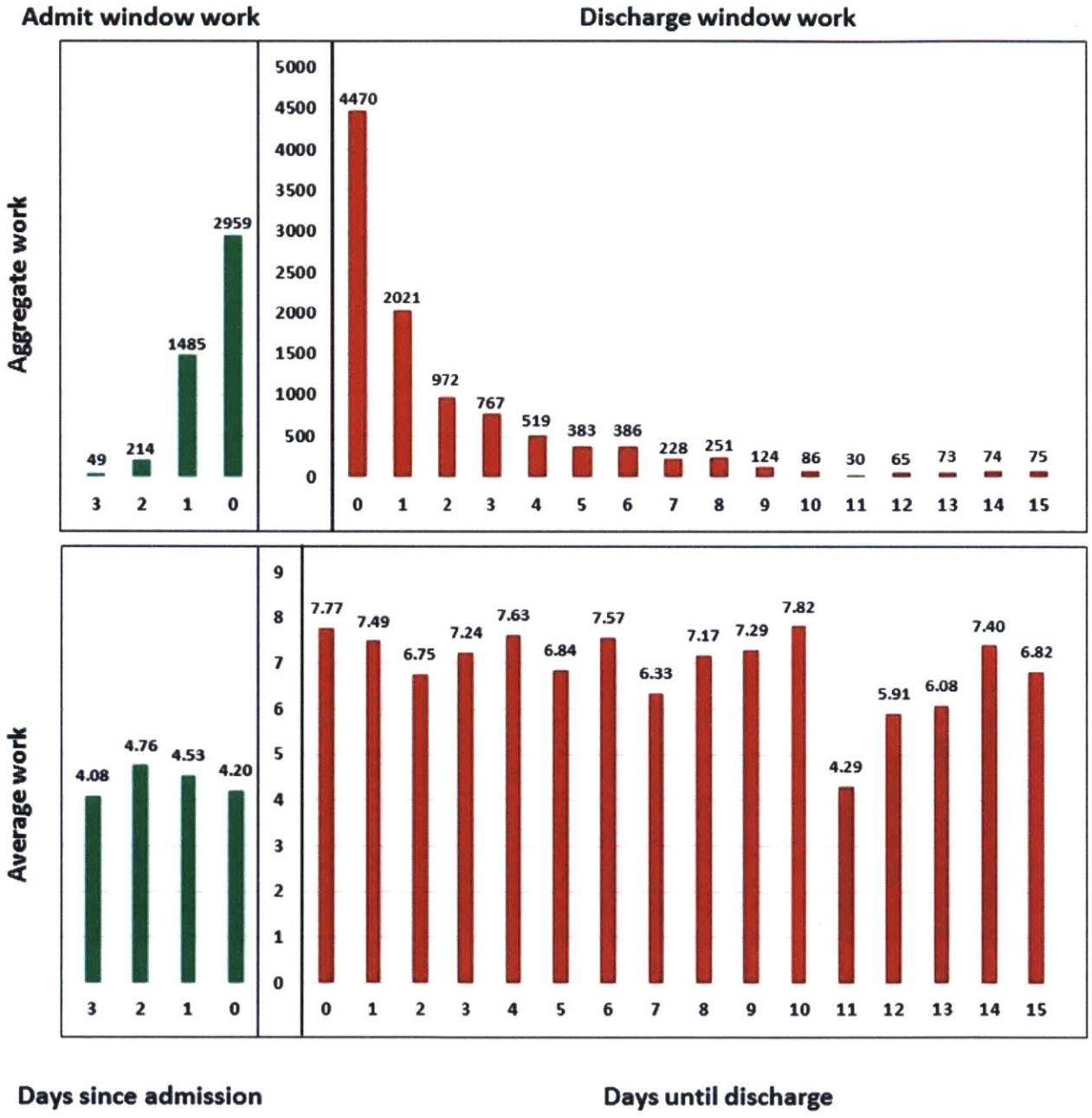
The rationale for a work metric rather than using note count as a proxy for workload was provided in Chapter 4 (the metric, though not completely refined, is more granular than note count), but, as should be expected, the distribution for notes, relative to admission and discharge, follows a similar distribution as shown in Figure 5-2.

Though Figure 5-1 may use a more granular metric, it is both far from comprehensive and provides an incomplete picture for both obvious and more subtle reasons. This chapter does not eschew statistical significance, but, despite its shortcomings, several aspects of Figure 5-1 can be discussed from the perspective of practical significance for a non-rigorous introductory discussion. First, the level of discharge window work, in the aggregate and on average, is greater than that of admit window work. This certainly does not hold for every patient as there are many patients, those not requiring case manager intervention, that would not have work in the discharge window. For example, of the 1229 patients considered for White 8, only 665 have discharge window work documented, while 934 have admit window work. The contribution of admit window work and discharge window work is considered more rigorously below.

Second, while the figure is truncated at 15 days until discharge and 3 days since admission, discharge window work can be documented further from discharge day than admit window work can be documented from admission. Again, using White 8 as an example, discharge window work is documented up to 90 days from discharge, while admit window work is documented up to 32 days since admission. Still the figure as presented captures 95% of the total discharge window work scored and 97% of the total admit window work scored. There is a third category of work, pre-discharge work, which accounts for a small component of the total work for White 8 and White 9. This work is typically documented in a weekly update note (WUP) and occurs for patients that have been assessed by the case manager but whose condition and post-MGH needs are so unclear as to preclude active discharge planning by the case manager. By our convention this is included in admit work (technically non-discharge window work). This work may warrant a separate category for floors with a very long LOS. This convention and others are explained more fully in section 5.3.

[1]This figure is shown over two pages, one for each floor.

White 8


Admit window work
Discharge window work
Aggregate work
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
49
214
1485
2959
3 2 1 0
4470
2021
972
767
519
383
386
228
251
124
86
30
65
73
74
75
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Average work
9
8
7
6
5
4
3
2
1
0
4.08
4.76
4.53
4.20
3 2 1 0
7.77
7.49
6.75
7.24
7.63
6.84
7.57
6.33
7.17
7.29
7.82
4.29
5.91
6.08
7.40
6.82
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Days since admission
Days until discharge

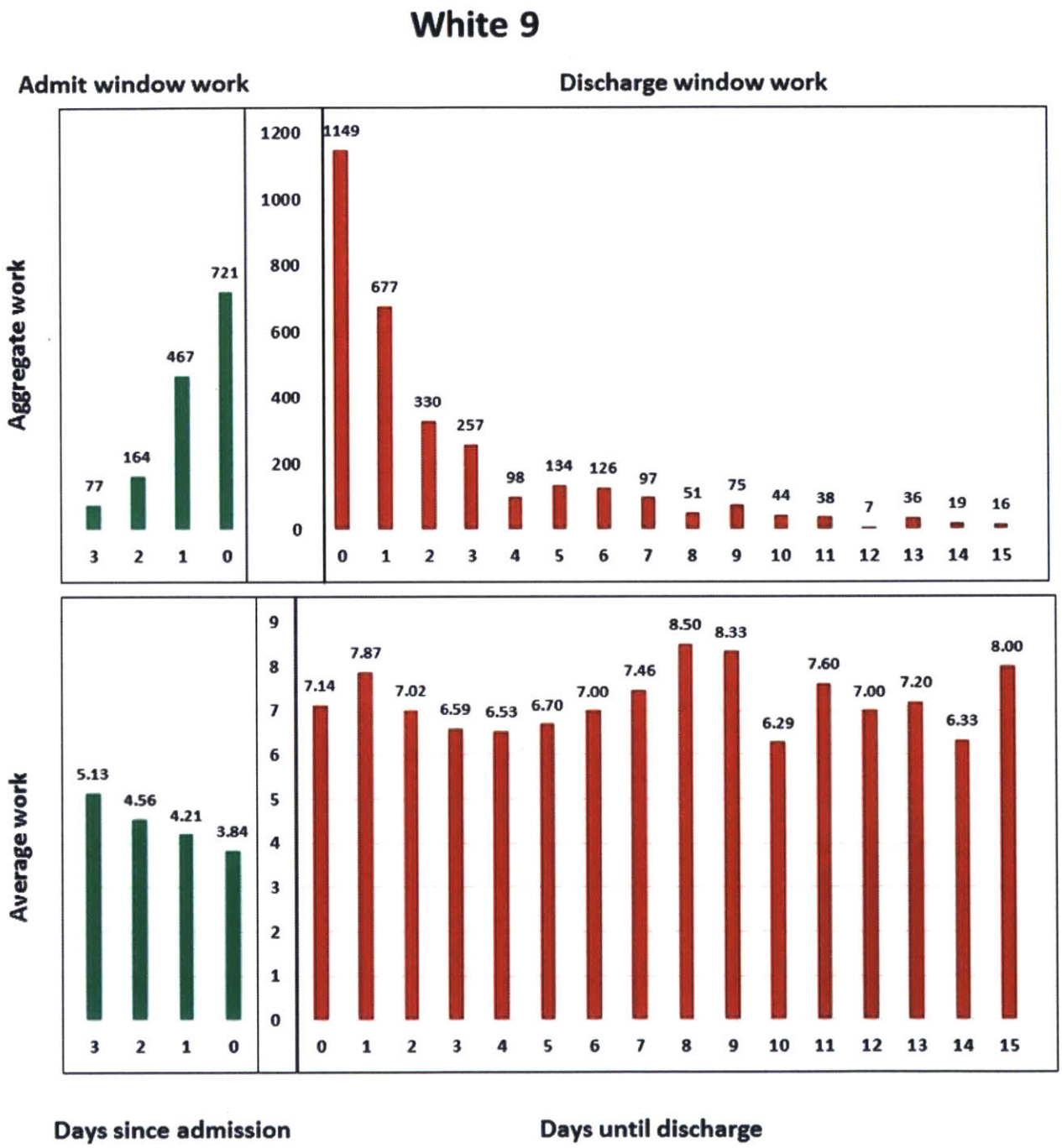


Figure 5-1: Aggregate and average distribution of case manager work relative to discharge and admission for White 8 and White 9, patient-level

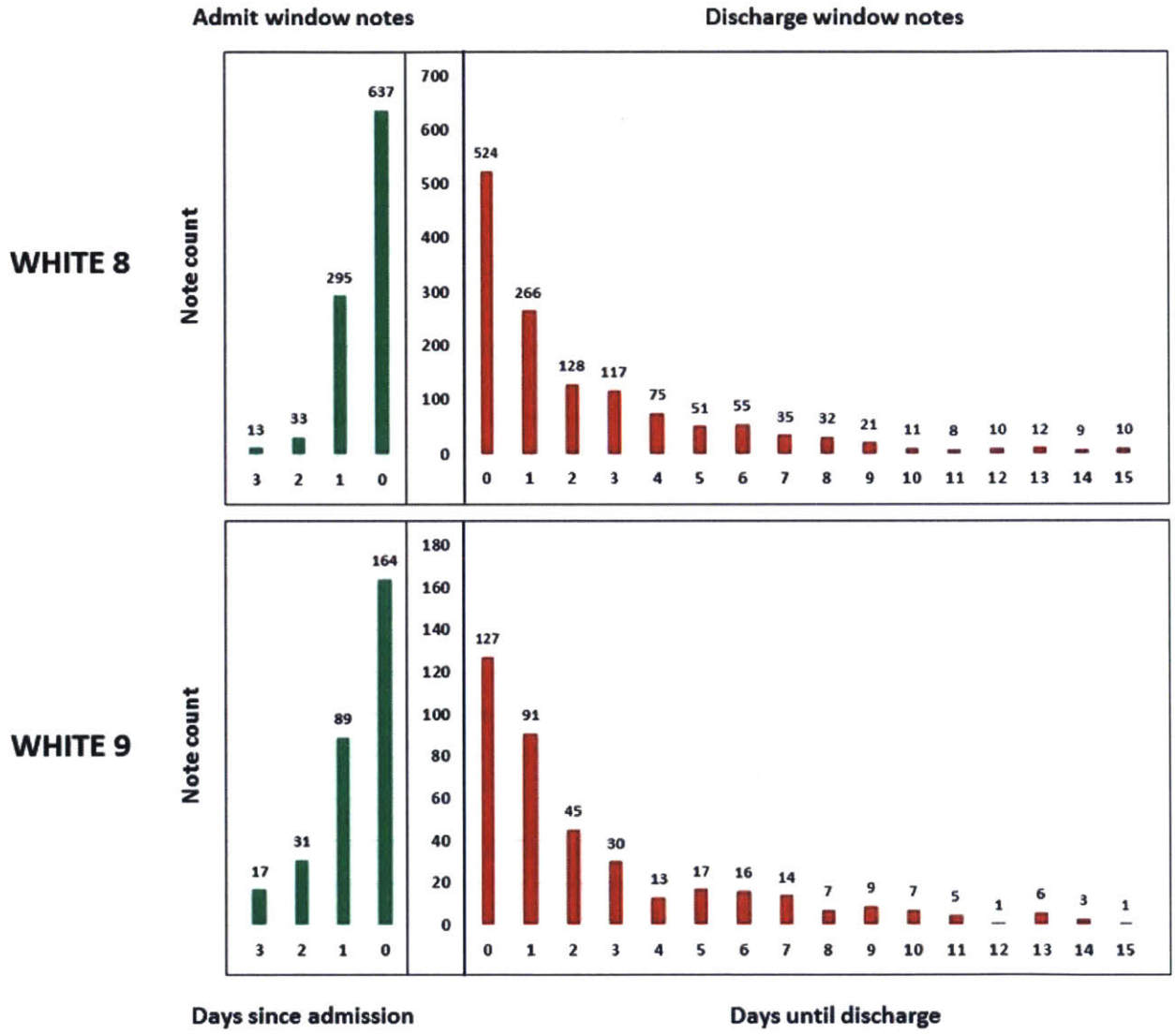


Figure 5-2: Distribution of case manager notes relative to admission and discharge for White 8 and White 9

Returning to a non-rigorous consideration of the figure, most of the admit window work occurs in the first few days of admission, with the plurality occurring on the day of admission. In another of our conventions detailed in section 5.3, the admission day is adjusted to account for late admissions to the floor and weekend/holiday admissions; i.e., since admit window work is not the focus of the reduced weekend staff a patient admitted on Friday after 1730 has her admission date adjusted to the following Monday for time calculations relative to admission. This adjustment is in line with MGH CM leadership's direction that the HR screen be performed within 24 business hours of admission and allows a more consistent comparison of work timing across patients.

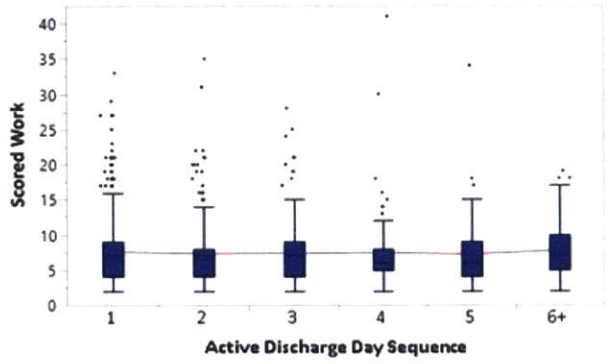
The distribution of admit work, both in the aggregate and on average, of White 8 compared to White 9 seems to hint at slight inter-floor differences in CM work patterns that were alluded to in Chapter 4. It is natural to ask if these are significant. In general, cross-floor statistical comparisons are problematic because of the relatively small sample size for White 9 and the correspondingly low power of appropriate tests. Some tests are presented below but, in fact, the observed differences are a consequence of slight differences in work patterns. Specifically, the initial admission work distribution is shifted slightly away from admission for White 9; i.e. the White 9 CM performs the HRIA slightly later, on average, than the White 8 CM. Thus admit work in one day since admission is a higher percentage of the one day since admission and admission day sum than for White 8. Additionally, the White 9 CM splits the HR screen and initial assessment (wbc-type notes from Chapter 4) more often than the White 8 CM. Since the HR is less work than the initial assessment this leads to an increasing average admit work as we move from admission for White 9. These effects and their consequences, along with other differences between the floors through the lens provided by our work metric are examined further in the penultimate section of this chapter.

Considering discharge work, it is tempting to draw conclusions such as most work is concentrated on discharge day and, regardless of the day, the average value of the work for a patient is the same. In fact, this is true but the figure is not constructed in a way to unambiguously support these conclusions. This is a subtle point but, because all discharge work for a patient is not on contiguous days, it is more illuminating to examine work going forward in sequence rather than back from the fixed, and largely unknowable, discharge day. Not to belabor the point, but a patient with two active discharge window days, for example, could have work performed on any two days. Not all of the work performed on the first active discharge day for these patients would be included in the aggregated work shown for one day until discharge and discharge day. The distribution of work in the discharge window by sequence of work episodes is considered below.

5.2 Successive refinement of analysis for discharge work distribution by sequence

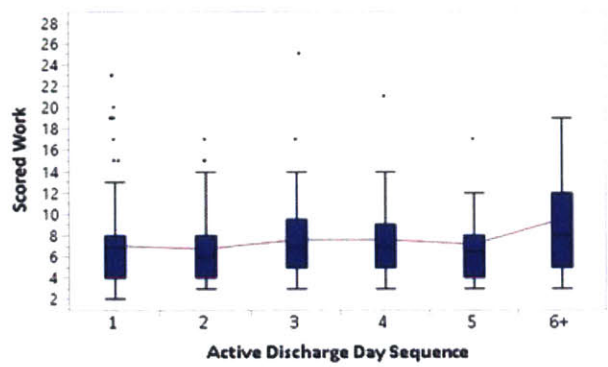
Consider a CM with a census of 20 patients(cases) on a given day. It is useful, if not entirely correct, to distinguish between active cases and inactive cases. Put simply, an active case is one that the CM is actively working on during that given day. There is some work, such as discussing a case at rounds, associated with every case (hence, the appellation active is not entirely accurate), but the majority of a CM's time for the day will be spent on work for active cases that necessitate phone calls, meetings, referrals, etc. Put another way, an active case is one where work events explicit or implied in a CM's notes can be scored and contribute to the workload metric. A case that is active today may or may not have been active yesterday and may or may not be active tomorrow,

even if the patient is still resident on the floor tomorrow. During a patient's time in the discharge window, active days are interspersed with inactive days. Figure 5-3 shows the discharge window work distribution by active day sequence for White 8 and White 9.



White 8

	Active Discharge Day					
	1	2	3	4	5	6+
n	655	368	186	111	57	110
Minimum	2	2	2	2	2	2
10th percentile	4	3	3	3	3	4
25th percentile	4	4	4	5	4	5
Median	7	7	7	6	6	7
75th percentile	9	8	9	8	9	10
90th percentile	13	12	12.3	12	12.6	12
Maximum	33	35	28	41	34	19
Mean	7.55	7.29	7.33	7.44	7.18	7.68
SD	4.29	4.43	4.41	4.97	5.12	3.62



White 9

	Active Discharge Day					
	1	2	3	4	5	6+
n	189	109	69	36	18	27
Minimum	2	3	3	3	3	3
10th percentile	4	3	4	3.7	3.9	3.8
25th percentile	4	4	5	5	4	5
Median	7	6	7	7	6.5	8
75th percentile	8	8	9.5	9	8	12
90th percentile	11	11	12	11.3	12.5	19
Maximum	23	17	25	21	17	28
Mean	7.01	6.70	7.54	7.56	7.11	9.41
SD	3.55	2.95	3.63	3.48	3.38	5.79

Figure 5-3: Distribution of discharge window work by active day sequence

Use nonparametric tests for means (Wilcoxon pairwise comparisons) or a median test reveals no support to conclude that any of the means or distributions for White 8 are different ($\alpha = 0.05$). For White 9 these same tests do support the alternative hypothesis that the day 6+ mean is different from day 2 and distributions differ. Possibly this is due to a small sample effects for day 6+ ($n = 27$), but even bearing this result in mind this basic analysis suggests that the amount of discharge work is similar regardless of the active day. This result can likely be traced to the point raised in Chapter 2 concerning high workload patients; namely, there are two basic types of high workload patients – those patients requiring highly idiosyncratic work ("unique" situations) and those patients requiring more "routine" work to be executed multiple times. Later active discharge days for these patients would contain a similar type and, thus, amount of work as earlier days. Of note, the variability as evidenced by the relative standard deviation (coefficient of variation) is substantial.

Whereas Figure 5-3 aggregates, for example, all first active discharge days regardless of the total number of active discharge days, Figure 5-4 shows the distribution of work by active discharge day for patients with exactly two active discharge days and exactly three active discharge days. On White 9, patients with exactly two active discharge days account for 28% of all patients with at least one active discharge day, while patients with three discharge days account for 17% of patients with an active discharge day(s). On White 9 these percentages are both 22%. Table 5.5 provides the frequency of patients with a given number of active discharge days for White 8 and White 9;

note in this table that the percentage of total value is the percent of all patients.

		White 8				White 9			
		n	median	mean	sd	n	median	mean	sd
Active discharge day	1 of 2	182	7	7.9	4.5	40	7	7.3	3.4
	2 of 2	182	7	6.9	3.9	40	6	6.6	2.3
	1 of 3+	75	8	8.2	4	27	6	7.1	4.3
	2 of 3+	75	6	6.8	3.8	27	6	6.7	3.6
	3 of 3+	75	7	6.8	2.6	27	8	7.7	2.6

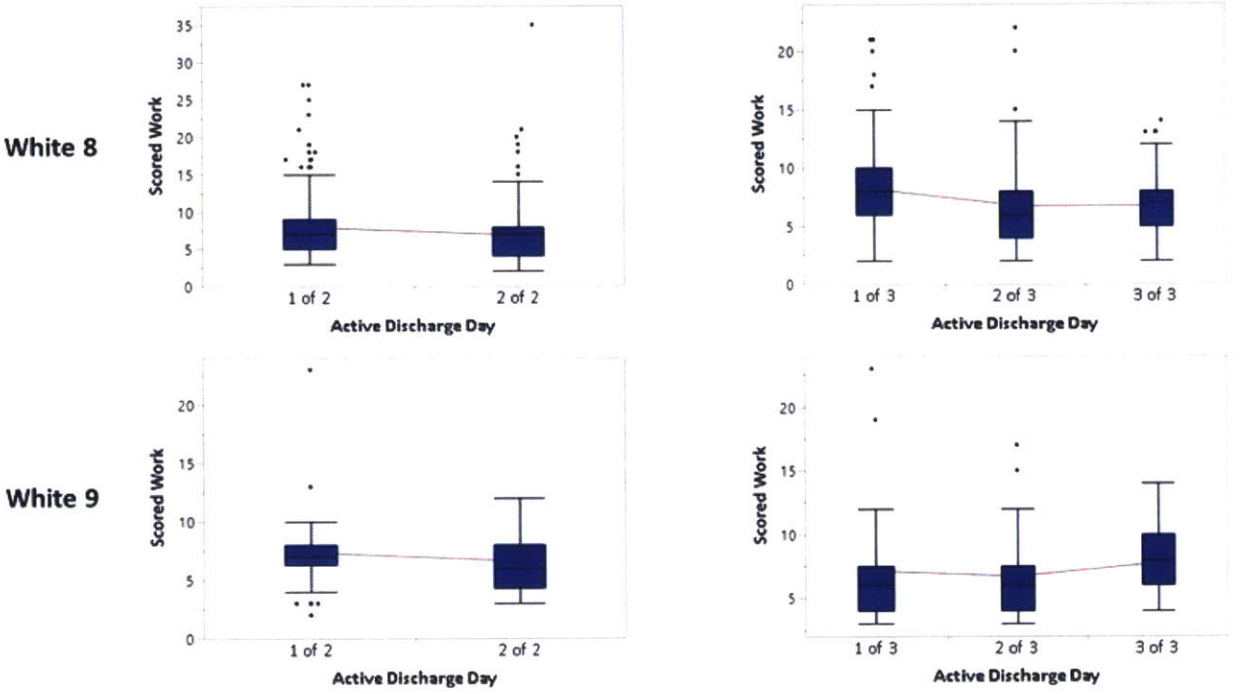


Figure 5-4: Summary statistics of discharge work across active discharge days for patients with two and three active discharge days

Similar to Figure 5-3, the importance of Figure 5-4 is in how similar the mean values for work are, in an absolute sense, regardless of where in the sequence (of active discharge days) the day occurs. Small sample sizes make statistical cross comparisons difficult, both within a floor between patients with a different number of discharge days, and across floors. With this caveat, nonparametric tests (Wilcoxon and pairwise Wilcoxon) suggest that, for patients with two active discharge days on White 8, the first day has statistically, though marginally, higher work. Looking at patients with three discharge days this pattern still holds. No such statistically significant ($\alpha = 0.05$) pattern is observed for White 9.

The key takeaway is that the amount of work done is similar regardless of active discharge day under a variety of aggregation schemes. Also, the variability is significant as measured by the standard deviation relative to the mean. This has consequences for daily workload prediction in that it is not sufficient, in general, to predict only the final discharge day for a patient with multiple active discharge days. Further, given the variability in active discharge day work scored, segmenting patients further by, for example, discharge disposition could lead to a lower variance predictive

Table 5.1: Count and frequency of patients with a given number of active discharge days

	White 8 patient total =1229		White 9 patient total = 392	
Active discharge days	n	% of total	n	% of total
0	574	46.7%	203	51.8%
1	287	23.4%	80	20.4%
2	182	14.8%	40	10.2%
3	75	6.1%	27	6.9%
4	54	4.4%	18	4.6%
5	23	1.9%	10	2.6%
6+	34	2.8%	14	3.6%

model. These issues are considered further in the following sections and Chapter 6.

Returning to an earlier point, measuring/predicting the timing of discharge work relative to the day of discharge is especially problematic prospectively. If one had the ability to do this then, implicitly, one is predicting the LOS for a patient, something that is notoriously difficult to do for a patient that will be discharging to a post-MGH facility. There are many issues, separate from a patient's clinical status, which contribute to this difficulty in predicting LOS. The consequences are that, for example, many patients with 3 or more active discharge days in the record could have discharged sooner (less active discharge days and less aggregate work) had complicating factors not been present. Therefore, a better perspective to take for daily predictive modeling would be to try and predict something easier than the final active day of discharge planning, such as the first active day of discharge planning. This prediction does not make any assumptions about the outcome of this first active day for predicting the current day's workload. Outcomes can be incorporated into modeling for subsequent days in fairly simple ways, again discussed in Chapter 6.

For the purposes of this chapter on current state analysis it is an open question whether the first active day of discharge planning is reliably predictable. The degree to which this day is predictable determines how exploitable the workload distribution across active discharge days shown below in Figure 5-5 may be.

The form of this figure is very similar to Figure 5-1, but this figure progresses by active discharge day rather than measuring discharge window work relative to discharge day. A large percentage of total discharge window work occurs during the first active day. Now the perspective afforded by the figure is skewed because of the large number of patients with a single active discharge day, as this is the first and last discharge day for these patients. Tables 5.2 and 5.3 supplement the figure and provide a more complete perspective.

The highlighted entries in Table 5.2 are mutually exclusive and sum to 100%, where the percentages are based on the sum of all discharge window work. In Table 5.3 the aggregate work by active discharge day for patients with varying numbers of total discharge days is provided. Here the percentages are based on column totals. Figures and tables for White 9, though not shown, follow a similar form.

The next section considers the daily distribution of work – admit window, discharge window, and

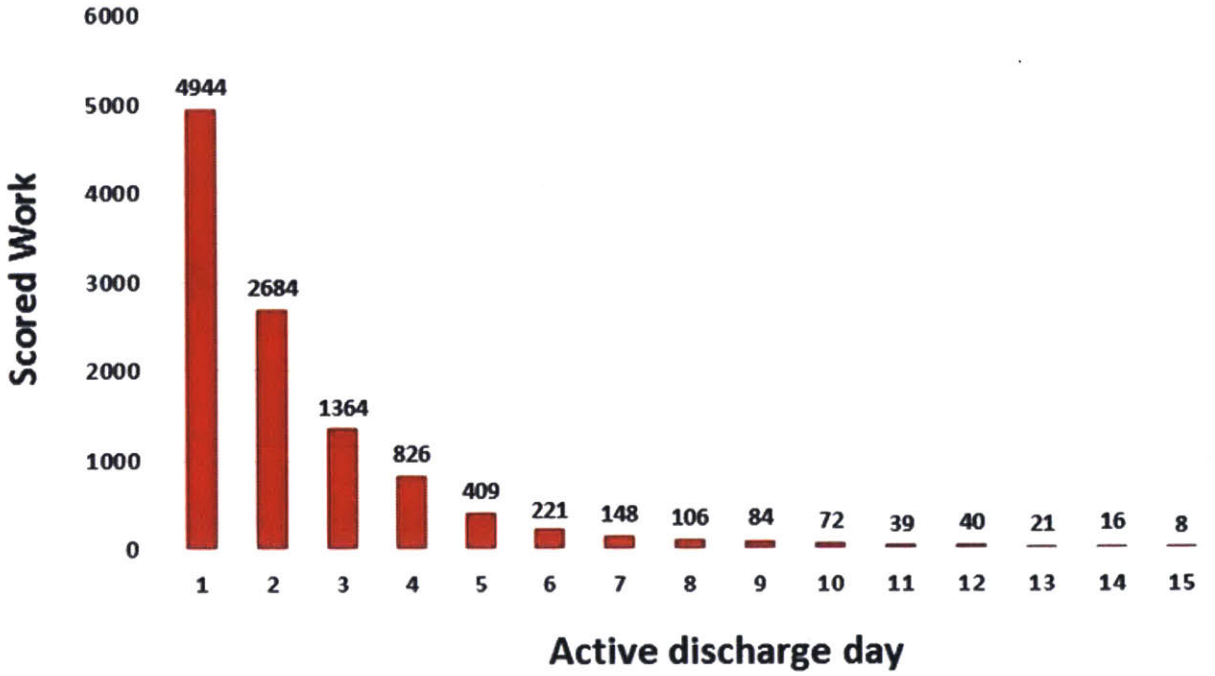


Figure 5-5: Distribution of discharge window work across active discharge days for 1229 patients on White 8, 1 October 2014 – 30 June 2015

Table 5.2: Contributions of different active discharge days to total discharge window work

	Scored discharge work	*% of total*
First active discharge day	**4944**	**45%**
First of multiple	*3101*	*28%*
First and only	*1843*	*17%*
Last active discharge day	**4457**	**41%**
Last of multiple multiple	*2614*	*24%*
Last and only	*1843*	*17%*
Intermediate active discharge days	**3436**	**31%**

Table 5.3: Allocation of discharge window work by active discharge day for patients with 2 – 6 active discharge days

		Number of active discharge days									
		2		*3*		*4*		*5*		*6*	
		Total	*%*	*Total*	*%*	*Total*	*%*	*Total*	*%*	*Total*	*%*
Active discharge day	*1*	1445	54%	612	38%	508	28%	209	23%	130	21%
	2	1254	46%	508	31%	477	26%	185	20%	106	17%
	3			508	31%	378	21%	204	22%	129	20%
	4					441	24%	168	18%	73	12%
	5							161	17%	94	15%
	6									99	16%
	Total	2699	100%	1628	100%	1804	100%	927	100%	631	100%

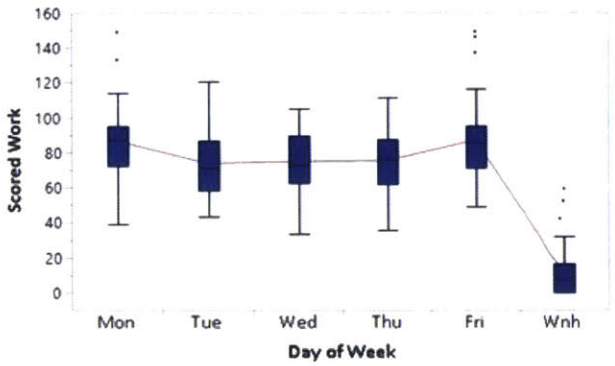
total – by day of the week. Though this is current state analysis, in many ways the pattern of this weekly distribution provides the key to beginning development of a predictive model. That is, explaining the observed weekday periodicity in terms of well-characterized quantities is necessary to have any chance of developing a predictive model of daily workload for case managers.

5.3 Daily distribution and weekly periodicity in case manager workload

The weekly periodicity of case manager workload was discussed in the opening sections of Chapter 4. At the time the proxy, high-level, indicator of case manager note count by day of the week was used to demonstrate this periodicity. Again, this periodicity can be traced to differential staffing patterns on weekdays vis-à-vis weekends and holidays. The limited number of case managers on the weekend primarily focus their efforts on facilitating the discharge of patients who were nearing discharge on Friday at the conclusion of regular business hours. Weekday case managers attempt to complete as much of the coordination for a pending weekend discharge as possible, often on Friday, in order to make the discharge facilitation function of the weekend case managers, who are effectively responsible, individually, for a larger number of cases across a greater number of floors, tenable. This can lead to more work for weekday case managers on Friday compared to the rest of the week. Additionally, because weekend case managers concentrate almost exclusively on discharge window work, the admit window work for patients admitted on the weekend, comprised of high-risk screens and initial assessments, is usually delayed until Monday, leading to more work on Monday as compared to other weekdays, sans Friday. Finally, there may be increased discharge window work on Monday as patients, not candidates for weekend discharge the preceding Friday, have progressed to the point where they are ready for active discharge planning, or discharge, on Monday.

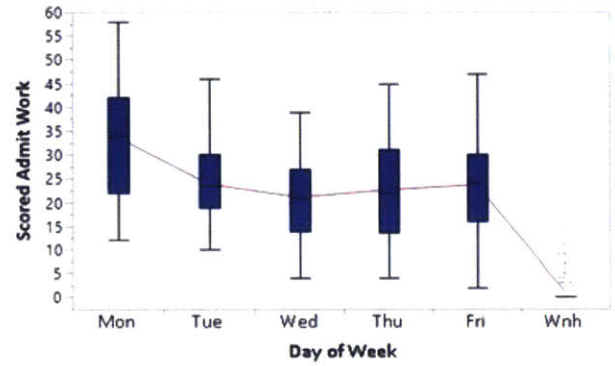
As stated previously, the weekly periodicity in CM workload is a MGH-wide, systemic phenomenon, reported by all case managers that we interviewed and presumably present on all floors. Figures 5-6 and 5-7 show the distribution of work by day of the week for White 8 and White 9 respectively. These figures use the score produced by the workload metric we developed and validated. The use of this metric allows the total work scored to be decomposed into admit window work and discharge window work components.

The picture painted in Figure 5-6 is compelling evidence to support the reported weekly periodicity of CM workload on White 8, as well as implicit further support for the workload metric if these reports are taken as a given. Considering total work, Mondays and Fridays are significantly "busier" than the other weekdays. Quantifying how much "busier" is difficult because the workload metric scale is not sufficiently calibrated but, if we consider the imputed time value of five minutes per unit of work score and the mean daily scores, 55-65 more minutes of work are documented on Friday and Monday compared to the other weekdays, on average. Using statistical, non-parametric pairwise Wilcoxon tests, the total work on Mondays and Fridays is statistically greater than on other weekdays, while the work on Friday and Monday is not statistically different, nor is the work on Tuesday, Wednesday, or Thursday. As mentioned in Chapter 4 (see footnotes), unadjusted pairwise tests may be problematic because with enough pairwise tests a meaningless statistical difference may seem apparent even at $\alpha = 0.05$. Our results include adjustments suggested for making multiple comparisons and associated statistical inferences.



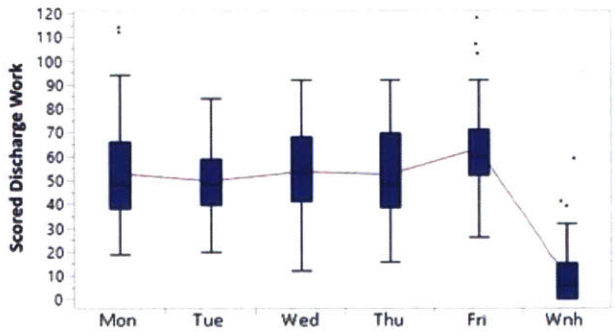
Total work scored

	Mon	Tue	Wed	Thu	Fri	Wnh
n	39	34	39	36	39	86
Minimum	39	43	33	35	49	0
10th percentile	63	44.5	50	48	58	0
25th percentile	72	57.75	62	61.75	71	0
Median	87	71	72	75.5	85	8
75th percentile	95	86.5	89	87	95	16
90th percentile	110	101.5	104	96.3	116	26.6
Maximum	149	120	105	111	149	59
Mean	86.38	73.44	74.54	75.00	86.74	10.93
SD	20.41	19.58	18.88	17.67	23.51	11.74



Admit work scored

	Mon	Tue	Wed	Thu	Fri	Wnh
n	39	34	39	36	39	86
Minimum	12	10	4	4	2	0
10th percentile	18	14	10	7.4	6	0
25th percentile	22	18.75	14	13.75	16	0
Median	34	23.5	21	22	24	0
75th percentile	42	30	27	31	30	0
90th percentile	48	31	33	40.2	41	5.3
Maximum	58	46	39	45	47	15
Mean	33.62	23.82	21.08	22.64	23.67	1.16
SD	11.40	7.57	8.81	11.56	11.83	2.90



Discharge work scored

	Mon	Tue	Wed	Thu	Fri	Wnh
n	39	34	39	36	39	86
Minimum	19	20	12	16	26	0
10th percentile	22	26	33	30.6	44	0
25th percentile	38	39.75	41	38.5	52	0
Median	48	48.5	53	48.5	60	6
75th percentile	66	58.75	68	69.75	71	15.25
90th percentile	76	73.5	77	82.6	92	25.3
Maximum	114	89	92	92	118	59
Mean	52.77	49.62	53.46	52.36	63.08	9.77
SD	22.81	16.17	18.21	18.79	19.46	11.26

Figure 5-6: Distribution of work by day of the week for White 8, 1 October 2014 – 30 June 2015

Looking at the second and third panels for the admit window and discharge window components of work, as hypothesized, admit work is highest on Monday with discharge window work highest on Friday. Using similar nonparametric statistical tests as described above, there is evidence at $\alpha = 0.001$ or lower to support the alternative hypothesis that the admit work on Monday is statistically greater than other weekdays (and, of course, weekends), while no statistical evidence emerges to suggest a difference among other days of the week. For discharge work, the mean level observed on Friday is statistically higher than observed for other weekdays, with no statistically significant differences between weekdays other than Friday.

The analysis, while providing independent support and verification of the reported weekly periodicity in CM workload certainly does not mean that Fridays or Mondays are always busier than other days. In fact, these days also exhibit the most variability in total work. This will be the basis for rudimentary examination of a pooling strategy to help reduce the daily variability in CM workload.

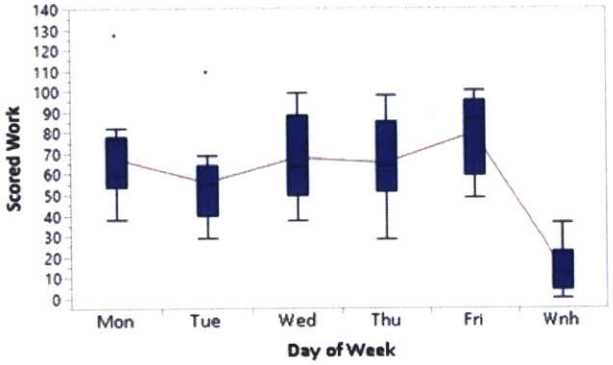
However, as a consequence of material presented in the next section, even a rudimentary analysis is not a straightforward exercise.

Figure 5-7, does not provide as much support for the reported weekly periodicity on White 9. This same weaker periodicity was also evident when using the high-level note count proxy in Chapter 4. For total work, the only statistically significant difference is between Friday and Thursday (α= 0.05). The relatively small sample size and correspondingly low power of statistical tests may account for why the level of work on Friday is not revealed as statistically different from the remaining weekdays, yet this does not account for the level of work scored for Monday, a level that was similar, over the examined time period, for Monday, Wednesday, and Thursday. Considering just the admit window component of work, the pattern observed is in line with reports; namely, the level of admit window work appears higher on Monday, even if nonparametric tests only reveal a statistically significant difference with Thursday. For discharge window work, Friday, in line with White 8 results and wider reporting by CMs, exhibits the highest level of discharge work, though the level only shows as statistically different from Monday at a 5% significance level.

The results for White 9 do not provide as strong of support for weekly periodicity in CM workload, or as much implicit support for the scoring of White 9 patients over this time period if this periodicity is taken as a feature at all times on all floors. Of course, with White 9 there were only thirteen data points available for each day, but the possibility of incorrect scoring has to be countenanced for several reasons. The initial text analysis, scoring, and validation was completed for White 8. The primary reason for this was based on consultation with CM leadership revealing the White 8 CM was notably thorough in documentation. This thoroughness was important because, even if a workload metric with perfect correspondence to reality was developed, if the record is incomplete then the score at the patient and day level is biased downward. As explained in Chapter 4 it is possible to infer and score work not in the record, but this process is not perfect – only the average value for a postulated missing note can be inferred. The amount of estimated missing work by day of the week for White 8 and White 9 is shown in Figure 5-8.

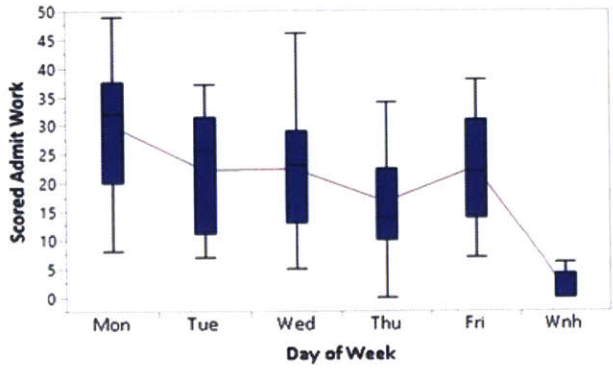
While not perfect, including the inferred work gives a work score at the patient and day level that is closer to the true value and it is the documented plus inferred work that is used in examining daily workload distribution and variability. In contrast, and as explained in Chapter 4, to test the validity of the text analytical techniques we developed to automate retrospective scoring, the documented work scores were used. In the majority of cases these scores were identical or very close (see Chapter 4). As discussed in the final section of this chapter there is evidence to suggest that the record for White 9 is not as complete as for White 8. This could lead to more error-prone scoring as more work has to be inferred. In addition, there is strong evidence to suggest that, over the time periods examined, the patient population for White 9 was significantly different from that of White 8 in a way that would decrease the amount of discharge work for White 9. This too is considered in the final section of this chapter. Finally, the periodicity for White 9, or rather the underlying causes of this periodicity may not have been as pronounced over the time period examined. This is considered in Chapter 6 by testing the performance of a predictive model of weekday CM workload developed with White 8 data on White 9. Explaining the weekday periodicity more rigorously, as well as explaining any observed differences between White 8 and White 9 is a powerful test for the model and its extensibility to other floors.[2]

[2]The need to have a generally applicable model versus models trained separately for different floors in considered in Chapters 6 and 7.



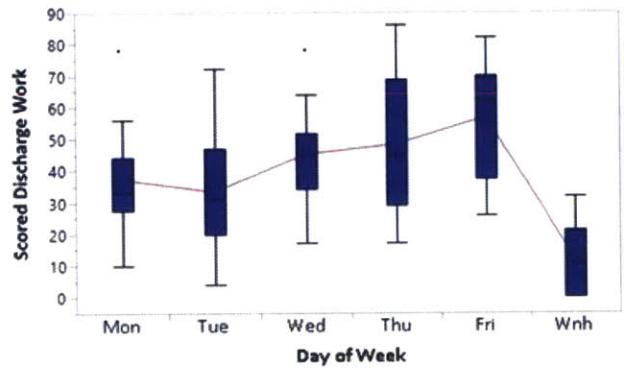
Total work scored

	Mon	Tue	Wed	Thu	Fri	Wnh
n	13	12	13	13	13	27
Minimum	38	29	37	28	48	0
10th percentile	43.2	31.4	38.2	32.4	48.4	0
25th percentile	53.5	39.5	49.5	51	59	4
Median	59	55	63	64	86	12
75th percentile	77.5	63.75	88	85	95	22
90th percentile	109	97	97.8	94	99.6	29.6
Maximum	127	109	99	98	100	36
Mean	67.08	55.50	67.38	64.77	79.23	13.59
SD	21.94	20.62	21.06	21.25	19.52	10.75



Admit work scored

	Mon	Tue	Wed	Thu	Fri	Wnh
n	13	12	13	13	13	27
Minimum	8	7	5	0	7	0
10th percentile	9.6	7.3	6.2	2.8	9	0
25th percentile	20	11	13	10	14	0
Median	32	25.5	23	14	22	0
75th percentile	37.5	31.5	29	22.5	31	4
90th percentile	48.6	35.8	41.6	31.6	36.8	5.2
Maximum	49	37	46	34	38	6
Mean	29.92	22.17	22.31	16.62	22.62	1.41
SD	12.51	10.36	11.24	9.23	9.47	2.26



Discharge work scored

	Mon	Tue	Wed	Thu	Fri	Wnh
n	13	12	13	13	13	27
Minimum	10	4	17	17	26	0
10th percentile	15.6	6.4	23	20.2	29.2	0
25th percentile	27.5	20	34	29	37.5	0
Median	33	31	45	45	62	11
75th percentile	44	46.75	51.5	68.5	70	21
90th percentile	69.2	66	72.4	80.8	78.4	29.6
Maximum	78	72	78	86	82	32
Mean	37.15	33.33	45.08	48.15	56.62	12.19
SD	16.73	18.69	15.36	22.06	17.38	10.50

Figure 5-7: Distribution of work by day of the week for White 9, 1 April 2015 – 30 June 2015

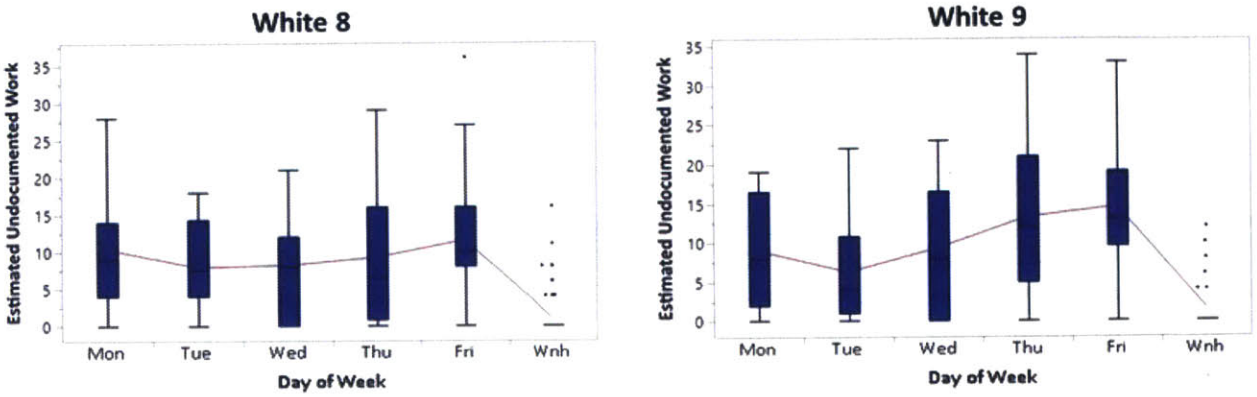
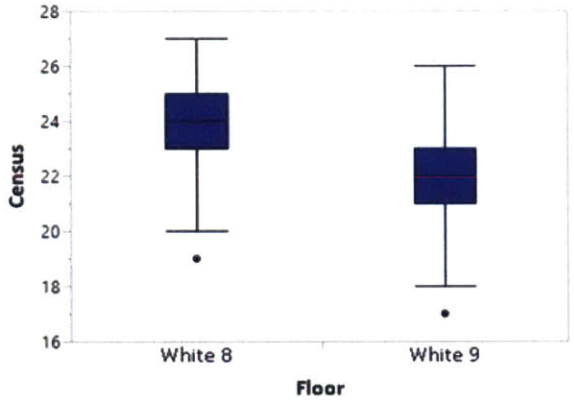


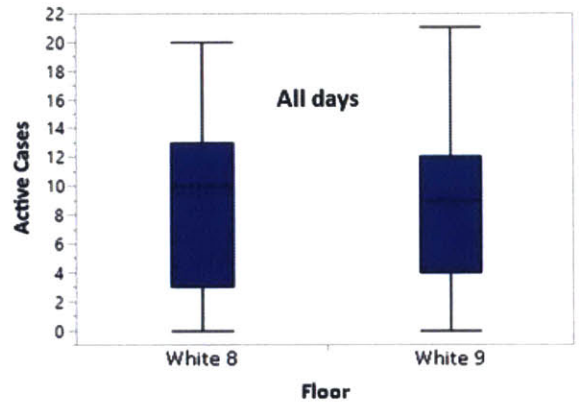
Figure 5-8: Estimated undocumented work for White 8 and White 9 by day of the week

The fundamental problem of observed workload distribution in a current state analysis, from a prediction standpoint, is explaining the observed variability; i.e. given a relatively constant census/case count, how can we explain the observed variability? From Chapter 4 we know that patients vary greatly in the amount of aggregate work they require, with the top decile of patients (by work score) accounting for 40% of the work over the time horizon considered. Yet, this fact, even if we could, with 100% accuracy, identify high workload patients, is not sufficient, and may not be necessary to predict daily workload with reasonable accuracy. Here reasonable means accurate enough to make operational improvements using the model directly or insights stemming from the model. Considering sufficiency, at the patient/case level, the fundamental source of day-to-day variability in workload is whether a CM is actively working on a case on a given day. From the daily perspective a patient with a high aggregate workload over the course of her LOS could represent "zero" workload on any particular day. It follows then that the fundamental source of variability in a CM's daily workload, aggregated across all patients, is the number of active cases. On one hand this is self-evident, but this fact assumes even greater prominence given the similarity, in absolute terms at the level of the mean, between the levels of discharge window work scored for a patient regardless of the active discharge day. Figures 5-9 and 5-10 illustrate the point concerning the fundamental impact of the number of active cases on daily workload and how workload tracks this number, as well as the variability in active cases as compared with the census.



White 8	
Sample	*Relative standard deviation*
Census	0.069
Active cases (all days)	0.619
Active cases (weekdays)	0.233

White 9	
Sample	*Relative standard deviation*
Census	0.084
Active cases (all days)	0.607
Active cases (weekdays)	0.313



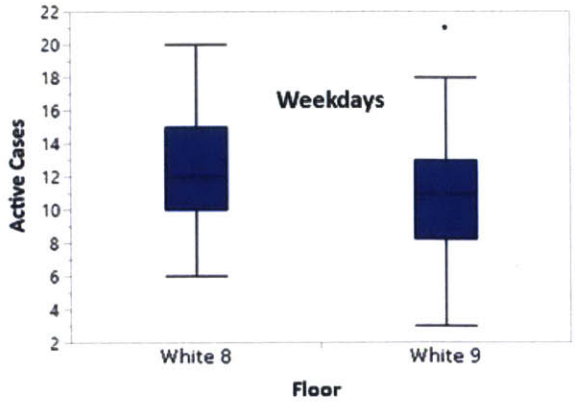


Figure 5-9: Comparison of census (caseload) variability with daily active cases variability

Of course, we are ultimately still left with trying to predict which cases will be active on any given day. The next section considers how to approach this task given a current state analysis considering

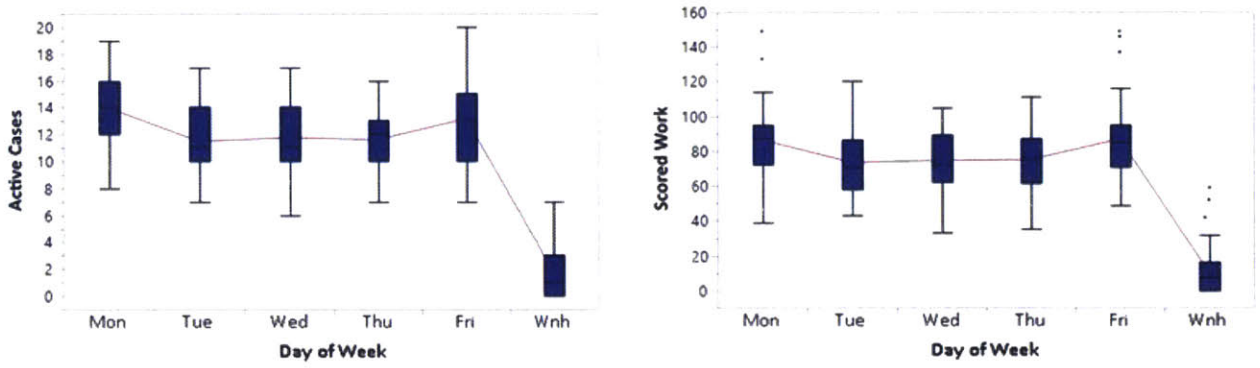


Figure 5-10: Comparison of the distribution of active cases and daily workload for White 8 by day of the week, 1 October 2014 – 30 June 2015

reference modes for work that may be defined for White 8 and White 9, the prevalence of the various modes, and how, at any point during a patient's stay, it is possible to identify the state that the patient is in. This will be a key building block of our daily predictive model as we seek to decouple modeling from a consideration of individual patient characteristics such as insurance, psych-social characteristics, or family dynamics, for example.

5.4 Common reference modes for case manager work

Previous sections have used terms like admit window and discharge window in an intuitive, non-rigorous way. This section formalizes the definitions of these terms and allows characterization of common reference modes for case managers. Additionally, this formalization allows generally unambiguous identification of the phase a case is in and, by extension, how many cases are in each phase for a case manager on a given day. This count by phase, in turn, is indicative of both what work needs to be completed for a case, and when this work will be completed.

Figure 5-11 is similar to a figure presented in Chapter 2 and describes how a patient progresses on the discharge planning plane during his LOS[3]. This is a generalized representation and common variations on this basic theme are presented after the figure is discussed. Table 5.4 provides the definition for the times indicated in the figure.

The patient is admitted to the floor at t_0 and is considered to be in the admit window (1) and in the unassessed state. After some interval (2) the patient is screened for high-risk (HR) criteria by the CM (3). This interval is supposed to be no more than 24 business hours from admission. The patient moves from the unassessed state to a partially assessed state and continues in this state until t_2 (4) when the initial assessment is completed for the patient. At this time the patient moves to the pre-discharge phase (5). This is generally a latent period for a case (6) when there is comparatively little activity. During this time the patient progresses on the treatment plane and eventually the patients probable needs upon discharge become sufficiently clear to allow active discharge planning

[3]This figure is fundamental for current and future predictive modeling work. The y-axis would correspond to our work metric (the foundation), while the figure connects the work done to the progression path for patients on the discharge planning plane; the connection provided by a phased framework is why Chapter 5 is referred to as the metaphorical bridge facilitating predictive modeling of case manager workload.

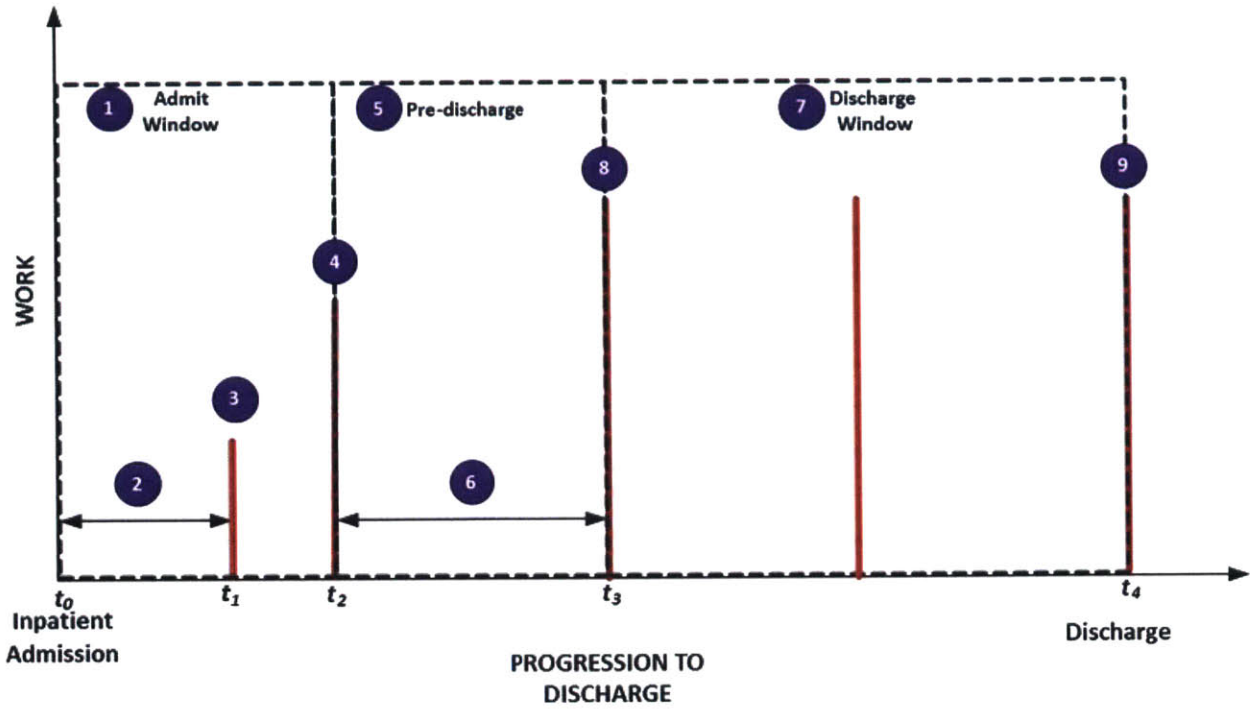


Figure 5-11: Patient (case) progression through discharge planning phases

at t_3 (8). At this point the patient enters the discharge window (8), generally a time of increased workload for the case. In the record this is typically indicated by notes of a specific type. During the discharge window the patient may have a varying number of active discharge days, which may not be contiguous. The patient continues in this phase until discharge at t_4 (9). In the figure workload events and there relative magnitudes are indicated by vertical red lines.

There are a few conventions to note in Table 5.4. First, as described previously, the admission time and discharge time to a floor is adjusted to account for weekends, holidays, late admissions, and early discharges. The adjusted floor LOS uses the adjusted admission and discharge times to calculate an inclusive measure. This is a convention used to facilitate current state analysis and maintain internal consistency of other conventions described below that allow consistent comparisons; this

Table 5.4: Definition of milestones during a patient's LOS

Time	Definition
t_0	Adjusted admission time to floor
t_1	Begin assessment
t_2	End assessment (end admit window) / Begin pre-discharge
t_3	End pre-discharge / Begin discharge window (phase of active discharge planning)
t_4	Adjusted discharge time from floor

Adjusted floor LOS = $t_4 - t_0 + 1$ (inclusive measure)

Table 5.5: Description of work reference modes and associated phase duration calculations

Reference Mode	Unassessed	Partially assessed	Pre-discharge	Discharge	Comments
1	$t_4 - t_0 + 1$	*Undefined*	*Undefined*	*Undefined*	Unassessed period is equal to adjusted floor LOS;"zero" work case
2	$t_1 - t_0$	*Undefined*	$t_4 - t_1 + 1$	*Undefined*	Only admit work
3	$t_1 - t_0$	*Undefined*	$t_3 - t_1$	$t_4 - t_3 + 1$ *	Admit work and one day of discharge work; typically one day of discharge work is on final day* for a one day discharge window
4	$t_1 - t_0$	*Undefined*	$t_3 - t_1$	$t_4 - t_3 + 1$	Admit work and 2 or more discharge work days
5	$t_3 - t_0$	*Undefined*	*Undefined*	$t_4 - t_3 + 1$	No formal admit work as discharge work, which may span a varying number of days, is first documented work
6	$t_3 - t_0$	*Undefined*	*Undefined*	$t_4 - t_3 + 1$	Formal admit work and first day of discharge work begin on same day ("combo" case); total active discharge days varies
7	$t_1 - t_0$	$t_2 - t_1$	$t_4 - t_2 + 1$	*Undefined*	Two or more days of admit work usually beginning with HR screen followed by initial assessment after one or more inactive days
8	$t_1 - t_0$	$t_2 - t_1$	$t_3 - t_2$	$t_4 - t_3 + 1$	Similar to reference mode 7 but admit work is followed by varying number of active discharge days
9	$t_1 - t_0$	$t_2 - t_1$	*Undefined*	$t_4 - t_2 + 1$	Similar to reference mode 8 but first day of discharge work occurs same day as final day of admit work

LOS differs from other values of LOS that may be calculated and its use is not meant to be extended beyond the scope of this work.

As explained, Figure 5-11 is highly generalized. Chapter 2 gave some idea of how the path for patients could vary based on the results of the initial HRIA. Figure 5-12 and the accompanying Table 5.5 show the common ways that a patient's progression on the discharge planning plane may vary in terms of reference modes evident in the case manager notes. These reference modes are for patients with White 8 (or White 9) as the only floor on which CM work was documented or implied. Seven additional reference modes, in addition to those in Table 5.5, exist to exhaustively cover transfer in, transfer out, or transit patients. These are not discussed, but, in general, these types of patients may enter the floor at any phase. The progression between phases is marked in the same manner as patients whose entire LOS is on White 8 or White 9. Similarly, the time spent in each phase is calculated in the same manner. The reference modes in Table 5.5 cover 922 of 1229 patients on White 8 and 299 of 392 patients on White 9.

Of course, there are many different ways to define reference modes for the observed CM workload, but, as described in Table 5.5, in the system presented the reference modes are distinguished by which phases a patient may be associated with, the number of discharge days, and whether admit work and discharge work occur on the same day. There is not always a clear-cut separation between admit work and discharge work but some distinguishers include whether a referral was placed or

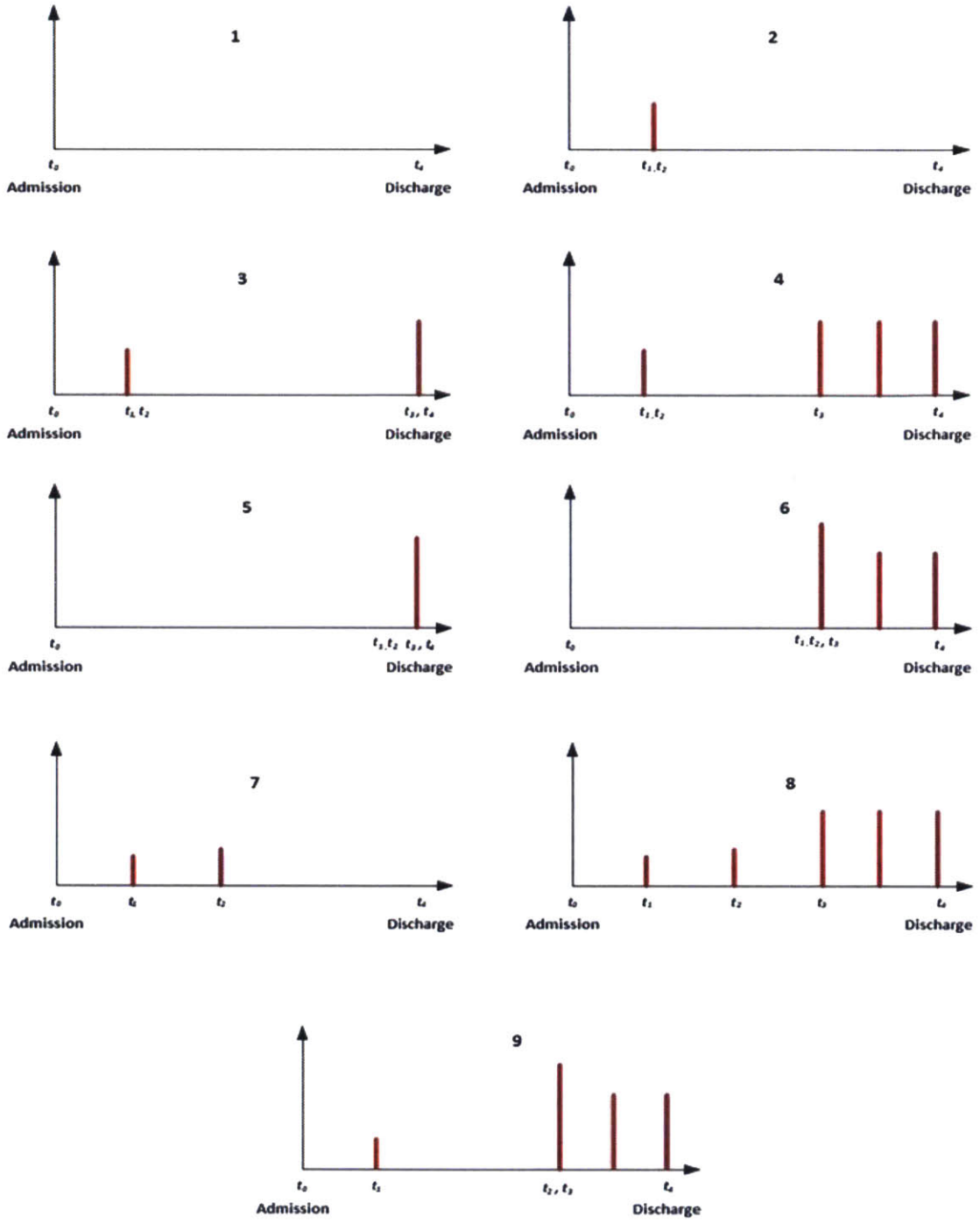


Figure 5-12: Reference mode for patients with all documented and implied work completed by White 8 and White 9 case managers

whether the work was primarily for assessment/ information-gathering purposes.

As textual markers were used to distinguish note types, so too can note types be used to distinguish which phase the work is occurring in. Typical associations between the note type and phase, as well as the point in a phase, in which the work documented in a note occurs, are given in Table

Table 5.6: Typical phases and times when note types occur

Note Type	Typical location (phase/sequence) in the record
acceptnote	Discharge window other than first active day
afa	Admit window, begin assessment
afaplus	Admit window, begin assessment
denynote	Discharge window other than first active day
dnmc	Admit window, begin assessment
dnmcplus	Admit window, begin assessment
ftxt	Various
homenote	Discharge window first day or last day
iaother	Admit window, begin assessment
ifscreen	Discharge window first active day
ifscreen2	Discharge window first active day
ifscreen3	Discharge window first active day
ift	Discharge window, second and final active day
iftplus	Discharge window, second and final active day
iftplus2	Discharge window, second and final active day
metnote	Discharge window first active day
nreadynote	Discharge window second active day
pnote	Admit window other than first active day, pre-discharge
psychcm	Admit window or discharge window, various days
refernote	Discharge window first or second active day
reviewnote	Various, often for transfer-in patient
spokenote	Discharge window first active day
treport	Discharge window first active day
v2	Admit window, first or second active day (final admit day)
v2ra	Admit window, first or second active day (final admit day)
vnaref	Discharge window first active day
wbc	Admit window first active day; usually followed by another admit day
wup	Pre-discharge

5.6. Typical in this sense means that work did not have to be repeated because of complicating case factors; e.g. a note type typically found on day 1 of the discharge window period could also be found on other days if work has to be repeated.

It would be disingenuous to not mention that the phase identification for a note, and the work documented in a note, is currently not as automated as other text analytical steps but, in reading the notes the phase can be determined with much greater certainty (essentially unambiguously) than manual scoring of work events. For most cases, the discharge window begins on the day of the first note following the initial assessment, unless the note is a weekly update note. The difficulty generally arises when admit work flows straight into discharge work, as for reference modes 6 and 9. It should also be made explicit that the transitions between phases, for any modeling purposes, are inter-day transitions.

Identifying different reference modes for work is useful for a variety of purposes. First, it allows easier tracking of how many cases are in each phase at the outset of any given day, a useful, though not complete, indicator of how much work may be required during the day for a CM. Second, when

considering how generalizable a predictive model using, in some way, a count of cases in each phase may be, it is necessary to determine the relevant prevalence of the reference modes between floors. Table 5.9 looks at the relative mode prevalence for White 8 and White 9. This table is important to an understanding of how work patterns may vary between floors.

From the table the most important points at the current level of discussion concern major differences between the two floors, particularly the disproportionate number of reference mode 1 cases ("zero" work cases) on White 9 and the predominance of reference mode 4 cases (1 active admit work day and 1 active discharge day) from the perspective of contribution to total work scored during the time periods examined. The different reference modes have an associated characteristic duration for the various phases as shown in the table.

Table 5.7 also introduces another important measure for each of the phases in which a patient may be located, active and inactive days. Table 5.7 aggregates and summarizes some of the information from Table 5.9 from the perspective of how long patients remain in each phase and the characteristic active / inactive days associated with each phase. The numbers presented are days as calculated using the conventions described above.

Table 5.7 illustrates more clearly what Table 5.9 shows and what has been mentioned in passing several times; namely, the admit work in White 9 is delayed slightly compared to White 8. This is evident in the longer amount of time, in the aggregate, patients spend in the unassessed state on White 9 compared to White 8. The practice of splitting the HR and initial assessment on White 9 is also more prevalent on White 8, as evidenced by the percentage of patient time spent in the partially assessed state. Noteworthy is the large amount of time spent in the essentially latent pre-discharge phase on floors. Generally the unassessed, partially assessed, and pre-discharge periods can be considered low workload periods for a case, relative to the discharge window, typically containing only one day of work during which the HRIA is completed. The occurrence of this work is generally easy to predict as it typically occurs within the first 24 hours of admission, either the adjusted admission date or the following day.

Also of note is the sum of the low workload period percentages for White 8 and White 9 and the workload percentages for the more intense discharge window. These are similar and are suggestive of the fact that some type of activity ratio, particularly for the discharge window, could be a basis for creating a conversion factor between different patient populations in an effort to establish baseline staffing levels across very diverse floors. In a practical sense it does not matter which of a number of possible patient characteristics lead to more active discharge days; a floor with a higher activity ratio requires more work, on average, per patient. This idea is considered further in Chapter 7.

The overall activity ratio (active days to total days) is given as a percentage in the "Total active" column. Similarly, two activity ratios for the discharge window are highlighted in yellow. The first is calculated by the total active discharge days divided by the total days. The second is calculated by subtracting the number of patients with a single discharge day from both the numerator and denominator, as these patients inflate the activity ratio. Notwithstanding the fact that active discharge days are not necessarily contiguous, the higher activity ratio of cases in the discharge window is a useful observation when modeling, particularly when augmented by the fact that a discharge window case active the preceding day is even more likely to be active the following day (see Chapter 6). Table 5.8 shows values for the activity ratios when considering all patients, not just the ones that spend their entire LOS on White 8 and White 9.

Table 5.7: Aggregate duration of patient stay (days) in each discharge planning phase for White 8 and White 9

		Unassessed	Partially assessed	Pre-discharge	Discharge window	Total	Total active	Total inactive	Total active discharge
White 8	days	492	45	2725	1818	5080	1965	3115	1126
n = 922	%	9.7%	0.9%	53.6%	35.8%		38.7%	61.3%	61.9% | 56.1%
White 9	days	262	65	764	532	1623	644	979	345
n = 298	%	16.1%	4.0%	47.1%	32.8%		39.7%	60.3%	64.8% | 59.9%

Table 5.8: Activity ratios for all patients on White 8 and White 9

	White 8 n=1229	White 9 n=392
Total	6887	2252
Active	*2494*	*817*
Inactive	*4393*	*1435*
Activity ratio	***36.2% | 33.4%***	***36.3% | 33.9%***
Total Discharge	2616	744
Active	*1487*	*448*
Inactive	*1129*	*296*
Activity ratio	***56.8% | 51.5%***	***60.2% | 55.4%***

Table 5.9: Summary statistics for reference modes of work on White 8 and White 9

	Reference Mode	n	% patients	% scored work	Unassessed (days)				Partially Assessed (days)				Pre-discharge (days)				Discharge (days)				Active days				Inactive days			
					Mode	Median	Range	Avg	Mode	Median	Range	Avg	Mode	Median	Range	Avg	Mode	Median	Range	Avg	Mode	Median	Range	Avg	Mode	Median	Range	Avg
W8	1	43	4.7%	0.0%	1	1	(1,8)	1.65													0	0	(0,0)	0.00	1	1	(1,8)	1.65
W9		31	10.4%	0.0%	1	2	(1,15)	2.48													0	0	(0,0)	0.00	1	2	(1,15)	2.48
W8	2	350	38.0%	10.3%	0	0	(0,4)	0.45					2	3	(1,42)	3.37					1	1	(1,4)	1.02	1	2	(0,37)	2.86
W9		117	39.3%	12.0%	0	0	(0,4)	0.56					2	3	(1,19)	3.53					1	1	(1,3)	1.05	2	2	(0,17)	3.04
W8	3	199	21.6%	18.1%	0	0	(0,5)	0.53					2	3	(0,15)	3.31	1	1	(1,4)	1.22	2	2	(1,4)	2.01	1	2	(0,15)	3.05
W9		42	14.1%	12.1%	0	0	(0,3)	0.60					1	2	(1,14)	3.21	1	1	(1,9)	1.29	2	2	(2,3)	2.12	1	2	(0,12)	2.98
W8	4	207	22.5%	50.1%	0	0	(0,4)	0.43					1	3	(1,22)	3.80	2	4	(1,43)	5.18	3	3	(14,14)	4.20	3	4	(0,37)	5.22
W9		40	13.4%	30.3%	0	0	(0,2)	0.30					2	4	(2,15)	4.78	2	4	(2,15)	4.78	3	4	(3,9)	4.08	2	3	(0,24)	4.50
W8	5	3	0.3%	0.4%	Undef	3	(0,4)	2.33										3	(1,4)	2.67	2	2	(2,2)	1.67	4	4	(0,4)	2.67
W9		6	2.0%	2.0%	0	0.5	(0,8)	1.83									2	2	(1,10)	3.67	1	1.5	(1,5)	2.33	0	1.5	(0,13)	3.00
W8	6	92	10.0%	13.8%	0	0	(0,4)	0.55									2	2	(1,23)	3.58	1	2	(1,13)	2.14	0	1	(0,21)	1.99
W9		35	11.7%	20.1%	0	1	(0,7)	1.80									1	2	(1,19)	4.23	2	3	(1,10)	3.23	0	2	(0,18)	2.94
W8	7	4	0.4%	0.3%	0	0	(0,0)	0.00	1	1	(1,2)	1.25	3	3	(3,3)	3.00					2	2	(2,3)	2.25	2	2	(2,2)	2.00
W9		2	0.7%	1.3%	Undef	0.5	(0,1)	0.50	Undef	3.5	(2,5)	3.50	Undef	3.5	(0,7)	3.50					Undef	4	(2,6)	4.00	Undef	4	(1,7)	4.00
W8	8	23	2.5%	6.9%	0	0	(0,3)	0.43	1	1	(1,8)	1.70	3	3	(1,8)	3.78	1	3	(1,49)	6.87	3	4	(3,16)	5.48	2	6	(0,37)	7.30
W9		12	4.0%	14.6%	0	0	(0,2)	0.33	1	2	(1,4)	2.08	2	2	(1,42)	5.75	1	3.5	(1,32)	6.08	5	5	(3,23)	6.58	5	3.5	(0,55)	7.67
W8	9	1	0.1%	0.2%	1	1	(1,1)	1.00	1	1	(1,1)	1.00					2	2	(2,2)	2.00	3	3	(3,3)	3.00	1	1	(1,1)	1.00
W9		13	4.4%	7.6%	0	0	(0,1)	0.23	2	2	(1,7)	2.54					2	2	(1,8)	3.38	3	4	(3,6)	4.23	1	1	(0,8)	2.00

Despite the effort to standardize calculations associated with the work reference mode framework, these definitions likely lead to slightly lower discharge window activity ratios for White 8 as compared to White 9. Since the sum of the duration of all phases is constructed to equal the adjusted LOS, an inclusive measure, the final phase a patient enters was used to maintain this consistency. Thus, the inclusive +1 in the adjusted floor LOS formula shows up in the final phase duration calculation. A plurality of patients on White 8 are reference mode 4 patients ending with a single active discharge day. While the overall activity ratios for White 8 and White 9 are very similar (overall ratios would be insensitive to which phase duration calculation is used to ensure the sum of all phase durations equals the adjusted floor LOS), the preponderance of reference mode 4 patients may artificially lower the discharge window activity ratio for White 8.

While some differences between the current state analysis for White 8 and White 9 are expected due to sampling, even given the ostensibly very similar patient populations, other discrepancies are potentially more troubling. The main discrepancy is the total amount of work scored for White 9, which seems very low, when scaled appropriately, to that calculated for White 8. This seeming discrepancy is considered more fully in the next, final section of this chapter before predictive modeling is examined in Chapter 6.

5.5 Examining the differences in total measured workload between White 8 and White 9

Since the White 8 and White 9 samples cover different time periods and a different number of patients, some scaling factor is needed to compare the total amounts of work scored for the two floors. A comprehensive scaling factor would need to account for differences in patient populations. Of course, if it was possible to do this then it would already be within our abilities to predict how much work a case with a given set of characteristics would take. As a first approximation we could consider the ratio of the number of patient records scored, 1229/392 or 3.135 when converting from White 9 to White 8, or the inverse when converting from White 8 to White 9. Another reasonable approximation would be to scale by the number of days and the capacity (number of beds) on White 8 and White 9 giving a scaling factor of 3.12. Yet another possibility would be scaling by the number of days and the average census, yielding a factor closer to 3.2 depending on what time of day the census is taken. Using 3.13 as our factor and considering White 8 as our baseline, the scored and expected values for key White 9 metrics are shown in Table 5.10, as well as the percentage by which the scored values differ from the expected.

Some of the "%difference" values in Table 5.10 are concerning upon initial consideration. These values are made even more troubling by a seeming lack of pattern to the sign of the differences. However, upon a more granular examination of the current state over the time periods considered, the values make sense. Even though the discrepancies are explainable, the fact that such differences exist, even for similar patient populations, must be part of the context when considering how generalizable the model presented in the next chapter may be.

In explaining the differences above we do have to control, if not for patient characteristics, at least for one key consequence of differences – discharge disposition. A true *ceteris parabus* comparison between even White 8 and White 9 is problematic because, frankly, when considering cases "all other things being equal" does not apply. The possible exception to this rule, from a case manager

Table 5.10: Comparing scored and expected values for key White 9 metrics using White 8 values as a baseline

	White 8	White 9		
	scored	scored	"expected"	% difference
Admit Work	4781	1493	1525	-2%
Discharge Work	11015	3160	3513	-10%
Implied Work	1853	718	591	21%
Admit Work Instances	992	438	316	38%
Discharge Work Instances	1480	312	472	-34%
"Zero" Work Cases*	43	31	14	126%

perspective, concerns patients that return home with no services upon discharge with no services. Table 5.11 lists the frequency of various discharge dispositions for cases on White 8 and White 9 during the time periods under consideration. The patients considered are only those that had discharged by 30 June 2015, and for which a discharge disposition was clear. This last point is not trivial as, though the discharge disposition is recorded in several databases, these databases often do not agree. In the cases of disagreement the discharge orders and case manager notes were consulted. In the table the key categories to consider are highlighted in yellow.

	White 8 1177 patients		White 9 384 patients	
	n	% of total	n	% of total
Transfer Out	81	6.88	25	6.51
Transfer In	152	12.91	52	13.54
Transit	15	1.27	7	1.82
Home - No services	472	40.10	182	47.40
Home - Services	272	23.11	71	18.49
Home - Hi-Tech Services	12	1.02	9	2.34
Return to admitting facility	45	3.82	8	2.08
Short-term SNF	124	10.54	47	12.24
Rehab	28	2.38	4	1.04
Long-term acute care	32	2.72	9	2.34
Medical respite unit - homeless	13	1.10	2	0.52
Hospice - Inpatient	6	0.51	3	0.78
Hospice - Home	18	1.53	3	0.78
Acute inpatient - Medical	4	0.34	2	0.52
Acute inpatient - Psychiatric	25	2.12	6	1.56
Expired	17	1.44	5	1.30

Table 5.11: Discharge dispositions for White 8 and White 9 patients

Performing a two sample test for proportions the categories in yellow indicate a statistically significant difference using Pearson's Chi-square test. Focusing on patients that required no services upon discharge the test yields a p-value of .0106 leading to a rejection of the null hypothesis that the

proportion of patients discharging home with no services is the same between White 8 and White 9. The p-values for similar tests on patients discharging home with services and home with hi-tech services are .0491 and .0462 respectively. Statistically, these are the only significant results with α= 0.05. The practical significance for proportion of patients discharging home without services is more important. This means that proportionally more patients on White 9 required no discharge work, work that is generally of a higher amount than admit work. If this result is considered along with the disproportionately high number "zero" work cases on White 9 shown in Table 5.20 (126% more than expected) the lower than expected work scored for White 9 is consistent.

Though consistent, the analysis indicates, at least on the scale of intermediate periods of time (on the order of three months), even similar floors could have populations that vary greatly in terms of work required. This is further support for the idea that a dynamic element of staffing or case assignment is needed even if appropriate benchmark caseloads can be developed. Similar floors like White 8 and White 9 would likely have the same benchmark and an unequal distribution of work would result. These periods could lead to periods where some predictive model exhibits differential levels of performance on "similar" floors.

The sign difference between admit work and admit work instances is also explainable, mainly by considering work pattern differences described earlier. As described, White 9 more frequently splits the admit work between two separate days (admit work instances), completing the HR on one day and the initial assessment on a subsequent day. The total work completed during these two sessions is essentially equivalent to the work completed in a combined HRIA session. Further, even the increased number of patients typically requiring no discharge work because they discharge home with no services would still usually have admit work (unless they are part of the "zero" work group in reference mode 1). Therefore, the admit work can be close to the expected while the instances of admit work (active admit window days) are significantly higher than expected for White 9. Note, these are supportable explanations though they are difficult to test. As alluded to in previous sections, the documented work for White 9 is not as complete for White 8, as evidenced by the higher than expected implied (estimated) missing work.

Using the current analysis of this chapter as a guide, with the caveats of this section in mind, Chapter 6 seeks to develop a linear model to predict the upcoming day's workload for case managers. This model will rely heavily on tracking the number of patients in each discharge planning phase described above.

Chapter 6

Predictive Modeling of Case Manager Workload

Predicting daily case manager workload involves two primary facets that, while not completely separable, may be considered in isolation to gage their relative importance. These two facets may be referred to, generally, as the timing of work and the magnitude of that work. While the magnitude is self-explanatory and, in this thesis, is reflected by the work score presented in Chapter 4, the timing aspect refers to whether a case is active, as the convention was defined in Chapter 5. This chapter demonstrates conclusively that correctly predicting the timing aspect of work is necessary to effectively predict a case manager's workload on a daily basis. In fact, it was an implicit acknowledgement of the preeminence of predicting the timing of work (is a case active?) rather than the exact composition of the work (is an active case an aggregate high workload case?) that led to development of the reference mode / discharge planning phase framework introduced in the previous chapter. In a demonstration of how timing and magnitude are intertwined at a fundamental level, predicting the timing of work also gives an indication of the magnitude of work associated with a case on a given day because the framework we use allows distinguishing between active discharge window cases and active admit window cases, the former of which is associated with more work on a daily basis compared to the latter.

In essence, the predictive model presented in this chapter relies on counts, at some point before the workday begins, of how many cases are in each discharge planning phase to predict the workload for the day. Augmenting these predictor variables are counts based on further segmentation of cases in the discharge window – discharge window cases active the previous day and discharge window cases not active the previous day. The importance of these counts as predictors of workload, as well as the basis for this importance, is considered in this chapter as representative of an exploitable correlation for predicting daily workload. To these counts may be added another variable, a count of cases that will enter the first active discharge day during the current day. The results of modeling with and without this "predicted first day of discharge (FDD)" are presented, as are suggested model improvements to obviate the need for overtly predicting FDD cases.

In fact, this chapter contains important discussion points concerning future improvements to the predictive model because, while implicit or explicit prediction of the active census (number of cases currently assigned that are active on any given day) is necessary to predict case manager workload

on a daily basis, this is not sufficient. The general recommendations set forth in this chapter provide a blueprint for improving the model, based on linear regression, to allow more accurate prediction of a daily workload score.

As an alternative to predicting a daily workload score, the framework predictor counts are used as the basis for a classifier, based on a boosted classification tree, to predict high, medium, or low workloads. The initial performance of the classifier is encouraging on validation sets, but this performance hinges on semi-supervised clustering to develop a stratified random partitioning scheme crucial to training an effective classifier. This scheme may be heavily biased to perform well on the current data set. Still, the same improvements suggested for a regression model used to predict a daily score will enhance the ability to predictively classify the upcoming day as high, medium, or low workload. Ultimately the classification problem may be more tractable, at least in the near-term, than the linear regression problem.

This chapter begins with a discussion of benchmark explanatory models to see what type of predictive performance is possible, what data (days) should be included in the model, and how performance measures, both relative (e.g., R^2, adjusted R^2, etc.) and absolute (e.g., RMSE) should be interpreted. **Section 6.1 uses explanatory variables that would not be known beforehand and, thus, are not meant to be the basis for prediction.** An important consideration of this section is the low correlation between the HW census and the daily workload.

The chapter then presents candidate predictive models, some of which could form the core of future improved models and some which suggest the nature of future improvements. **These models are presented in sections 6.2 and 6.3 and use only predictors that would be known at the time prediction of daily workload is made.**

The chapter concludes with a consideration of the classification problem. Model improvement suggestions are not confined to the second section but are presented throughout, *in situ*, if you will, so that the context for the suggestions remains clear. Though current state analysis is formally the domain of Chapter 5, it is impossible to dispense with current state analysis in this chapter. In fact, the act of modeling provides important insights into the current state. Finally, the focus of this chapter is more on White 8 than White 9, mainly because of the larger White 8 data set. Except for important distinctions that are noted, the characteristics cited for White 8 also hold for White 9.

6.1 Benchmark performance of *ex post facto* explanatory models[1]

Table 6.1 on the following page provides some candidate explanatory models that deserve brief consideration to provide relevant context for this chapter, in terms of prediction difficulty, what data should be excluded from consideration, and caveats associated with performance measures, particularly relative performance measures. The first three models are in consideration of the observed weekly periodicity in case manager workload, generally higher on Mondays and Fridays compared to other weekdays, and very low, as consequence of staffing patterns and staffing priorities, on weekends and holidays. The relative performance of this model ($R^2 = 0.77$), just using days of the week as explanatory variables, is very high, suggesting that weekly periodicity is an important

[1] Again, for benchmarking purposes all variables, including ones not knowable *a priori* are used in this section. In sections 6.2 and 6.3 only true predictor variables are used for candidate models.

Table 6.1: Benchmarking possible OLS model performance

Model	White 8 Data Set	Explanatory Variables	Response Variable	R^2	Adj R^2	RMSE	Mean Response
1	All days	Days of the week	Total work	0.77	0.66	17.90	57.90
2	All days	Days of the week	Admit work	0.65	0.64	8.80	17.50
3	All days	Days of the week	Discharge work	0.61	0.60	17.20	40.40
4	All days	Active cases	Total Work	0.91	0.90	11.40	57.90
5	All days	Active admit cases, active discharge cases	Total work	0.92	0.92	10.60	57.90
6	All days	Active admit cases	Admit work	0.90	0.90	4.52	17.50
7	All days	Active discharge cases	Discharge work	0.86	0.86	10.30	40.40
8	Weekdays, no holidays	Days of the week	Admit work	0.16	0.14	10.42	25.00
9	Weekdays, no holidays	Days of the week	Discharge work	0.06	0.04	19.29	54.40
10	Weekdays, no holidays	Active admit cases	Admit work	0.78	0.78	5.29	25.00
11	Weekdays, no holidays	Active discharge cases	Discharge work	0.63	0.63	12.00	54.40
12	Weekdays, no holidays	Active admit cases, active discharge cases	Total work	0.65	0.64	12.44	74.40
13	Weekdays, no holidays	Census	Total work	0.02	0.01	20.70	74.40
14	Weekdays, no holidays	High workload census, low workload census	Discharge work	0.07	0.06	19.10	54.40
15	Weekdays, no holidays	High workload discharge window, low workload discharge window	Discharge work	0.11	0.10	18.60	54.40
16	Weekdays, no holidays	High workload discharge window, low workload discharge window, first active day of discharge	Discharge work	0.33	0.30	16.10	54.40

variable to include or control for in any predictive model.

Of course, this line of thinking is fundamentally flawed, but examining these flaws is illuminating. First, the coefficient of determination is high only because of the low level and low variance of the weekend work score. Second, despite the high score of the model by a relative performance metric, the performance of the model using an absolute performance measure (RMSE) relative to the mean response shows how poor this model is; the RMSE for model 1 is 17.9 with a mean response of 57.9. Controlling for the day of the week does not explain anything, it merely reflects the fact that some days are historically associated with a higher workload. What the weekly periodicity reflects is that more cases are active on Monday and Friday as compared to other days. Thus, models 4-7, using the active census as explanatory variables, have a much better relative and absolute performance than comparable models 1-3. Yet, even here, the relative performance is inflated because of the low-level / low-variance work on the weekends. In fact, it is possible to develop a predictive model using counts of cases in the various discharge planning phases, controlling only for weekends/holidays, with an adjusted R^2 exceeding 0.75. Again, this model does not predict anything contrary to what a cursory consideration of the coefficient of determination would suggest.

Models 8-12, with weekends and holidays excluded, make this point more starkly. Comparing model 9 to model 3 (or model 8 to model 2), when holidays and weekends are excluded, the relative performance of the model has decreased by a factor of more than 10, but the absolute performance of the model, though still unacceptable, has actually increased. This suggests that weekday periodicity, while present, is not so prevalent as to preclude operational changes such as pooling. This is considered in Chapter 7 but, suffice it to say, if weekday periodicity existed on all floors at all times, pooling would not be a viable strategy.

Comparing models 10-12 with models 8 and 9 bears on three key points of this section. First, the active census "explains" 65% of the observed weekday work score variance. Second, scoring the admit window work and discharge window work together (total work) is more effective than scoring each component individually and summing to explain a day's total work. Third, and related to point 2, though not obvious in Table 6.1, while a prediction of discharge window work could be improved with model refinements that consider specific features of the active discharge window cases, there is a more pronounced upper limit on the performance of scoring admit window work. The magnitude of this work depends on the nature of the initial interaction between the case manager and a patient. If a patient does not meet high risk criteria this interaction may be brief or non-existent. If the patient meets high risk criteria but does not require case manager intervention, the initial interaction would also be briefer (less work) than the initial interaction with a patient meeting high-risk criteria and possibly requiring case manager intervention.

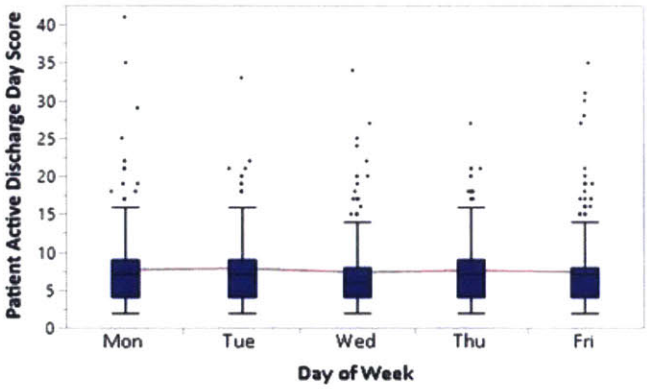
The point is that even if the number of active admit window cases could be predicted with certainty assigning anything more than an average work score for this work is not really supported. This does not need to be overstated but, as discussed in section 6.2, although discharge window work typically represents the largest component of a case manager's daily work this is not true on all days. For admit work-dominated days there is a significant level of currently irreducible uncertainty. Of course, one can envision a future automated pre-screening that gives a better indication to predict this admit window component of total daily work.

6.1.1 Examining whether the day of the week (weekday) should be considered in a predictive model

The model presented in this chapter does not consider weekends and holidays. This is in part because of issues discussed above. Furthermore, the count variables used for prediction are stripped of their information value on the weekends and holidays. As discussed in preceding chapters there is a greatly reduced level of case manager staffing on the weekend. These weekend case managers rely, in part, on a worklist that specifies discharge window cases that may be ready for discharge on the weekend to focus their efforts. The idea of the worklist is one that bears future examination because if this projected worklist is correlated with the active weekend census and workload then it reveals the extent to which an active discharge window case can be reliably forecast; the ability to accurately forecast the first, last, or any intermediate active discharge day on weekdays would be an important input to a predictive model.

Even focusing exclusively on weekdays, the question still remains as to whether the day of the week should be controlled for in a predictive model. On the one hand, any significance attached to a day of the week variable may merely reflect the historically observed pattern. However, there may be good reasons to control for the day of the week, particularly Friday and Monday. Consider, discharge work that may normally be completed on the weekend (if staffing levels were similar to Friday)[2] is often shifted to Friday in order to facilitate the work of weekend case managers effectively responsible for more cases, but with a narrower focus, than weekday case managers. Further, cases not projected for imminent discharge on Friday may have progressed to that point over the intervening weekend. This may account for more work on Monday. The question of whether and how to control for the weekday hinges on the answer to the question: Is the generally greater workload on Monday and Friday the result of more active cases and/or the result of more work done per active case?

Figure 6-1 results from attempting to provide an answer to this question. In this figure the amount of work per active discharge window case on weekdays is shown.



	Day of Week				
	Mon	Tue	Wed	Thu	Fri
n	264	219	283	249	333
Minimum	2	2	2	2	2
10th percentile	4	4	4	4	4
25th percentile	4	4	4	4	4
Median	7	7	6	7	7
75th percentile	9	9	8	9	8
90th percentile	13	13	12.6	14	12
Maximum	41	33	34	27	35
Mean	7.63	7.84	7.31	7.54	7.37
SD	4.89	4.40	4.32	4.09	4.49

Figure 6-1: Work per active discharge window case by weekday for White 8, 1 October 2014 – 30 June 2015 (weekends, holidays excluded)

Figure 6-1 (shown with connected means) supports several inferences. First, the work done per active discharge case, regardless of day, is similar on the whole; statistically the level is indistinguishable. However, looking at the number of active discharge cases for Friday (333) compared to other

[2]It is important to remember that sub-acute facility staffing levels are also reduced on the weekend.

weekdays does show there are significantly (both from a practical perspective and statistically, via a non-parametric test on the daily means) more active discharge cases on Friday. The increased Monday workload is, to a large extent, explained by the patients admitted over the weekend requiring a high-risk screen and initial assessment on Monday (delayed work).

Further analysis is needed to explain the generally higher Friday workload. If work typically completed on multiple days were moved forward then one would expect a significantly higher mean. The figure is not definitive, however, because of the distribution of implied work by weekday presented in Chapter 5. The most implied work for White 8 occurred on Fridays and Mondays. The possible reason for this observation, the low prioritization for full documentation when faced with a high workload, was considered. The practical significance is that there was often no basis for imputing more than an average value for implied work. Figure 6-2 considers only explicitly documented work.

The left panel of Figure 6-2 is analogous to Figure 6-1 with weekends included. When considering only explicit work the results are consistent with considering implied (undocumented) plus explicit (documented) work. Of note, the mean for work (per active case) on Monday is statistically higher than Friday and Wednesday (non-parametric Wilcoxon test on means at $\alpha = 0.5$), suggesting there may be a "catch-up" effect as weekday CMs work to facilitate the discharge of patients who have become ready for active discharge planning over the preceding weekend. This, however, is speculative rather than definitive.

The lower mean value for Friday again suggests that the driver of workload on Friday is the number of active cases rather than an increased amount of work completed per active case. The right panel of Figure 6-2 considers Fridays in more depth by looking at the documented work completed per active discharge case in three groups: patients discharging the same Friday, patients discharging over the same weekend, and patients discharging at some later time.

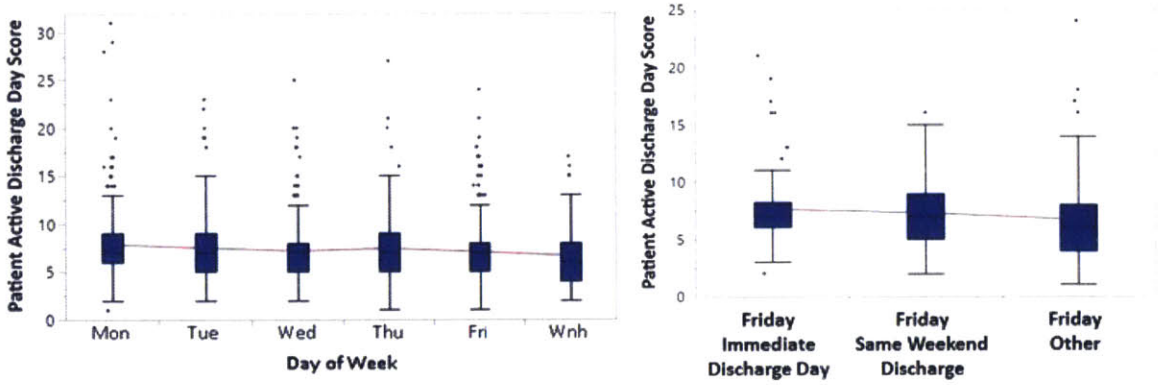


Figure 6-2: Work per active discharge window case by weekday for White 8, 1 October 2014 – 30 June 2015 (explicitly documented work only)

The documented work for patients discharging on Friday is statistically, though marginally, greater than the work completed on Friday for patients discharging at a later time, though statistically indistinguishable, at α=0.05 from work per active case completed on Friday for patients discharging on the weekend. This is further evidence to suggest the driver of workload on Friday is the number of active cases rather than increased work per active case. Of course, this begs the question as to why there are more active discharge window cases on Friday. The answer is "explained", in part,

by considering Figure 6-3. There are more cases with Friday as the first day of active discharge planning (FDD) and, considering Friday and weekend days as a group, there are more discharges on Fridays/Weekends. Taking the other days as a group, a Pearson's chi-squared test for proportions (significant below $\alpha = 0.0001$) suggests a statistically significant higher proportion of patients with a FDD of Friday or discharge on Friday/Weekend. Of course, the raw counts of what occurred over the time period examined is far more compelling than any statistical test.

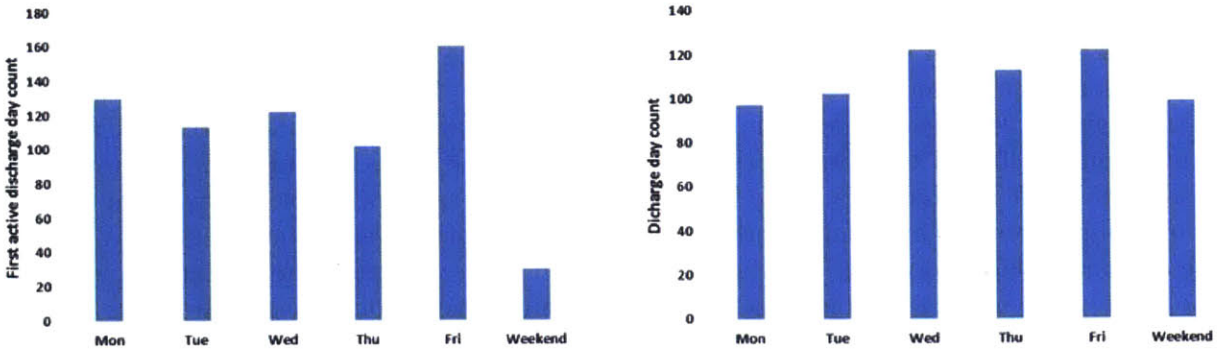


Figure 6-3: Count of first day of active discharge planning and discharge day-actual by day of the week White 8, 1 October 2014 – 30 June 2015

Figure 6-3 only considers patients with active discharge days; i.e. patients that enter the discharge window described in Chapter 5.

Returning to the question posed at the outset of this subsection, we chose not to control for day of the week in our predictive model of section 6.2, instead including predictor count variables that explained periodicity. Controlling for Friday does improve model performance to a slight degree, but this is indicative of the improvements that need to be made in future iterations of the model.

The fact that the average work for patients discharging on Friday suggests only one day of work is moved up from the weekend to Friday (more active cases rather than significantly more work completed per active case) has important implications for modeling and, more broadly, for any operational recommendations. Considering the latter, this is indicative of a request-response cycle delay. That is, to facilitate the discharge of a patient often requires communication with outside entities, such as post-MGH sub-acute facilities. The CM requires a response from the facility, often not immediate, that allows continuation of the current course of action, or requires plan alteration and a possible repeat of earlier planning steps. Thus, the work boundary we consider – the work completed by CMs and the timing of the work – is influenced by many entities outside of a CM's sphere of influence. From a modeling perspective this allows us to further segment the discharge window patient population/census into cases that were active the previous day and cases that were not. This segmentation is considered further in section 6.1.3, following a discussion of the inadequacy of knowing the population of aggregate high workload cases for predicting workload on a daily basis.

6.1.2 Examining the predictive value of a high aggregate workload patient count

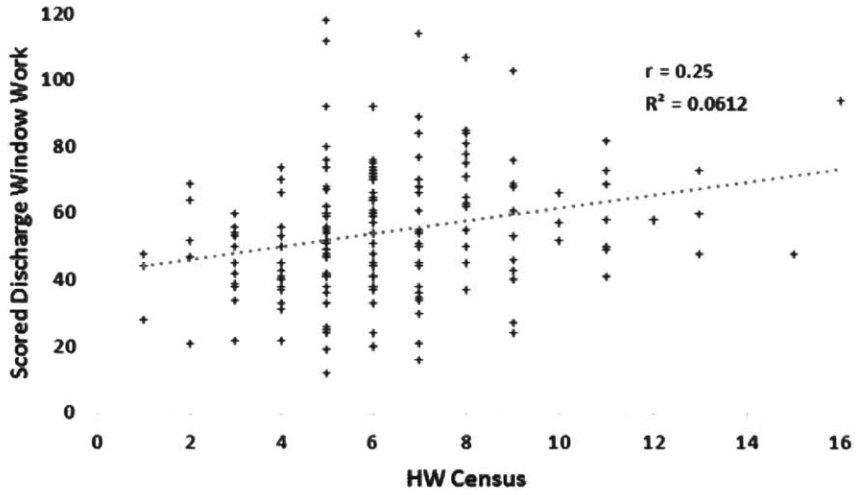
Returning to table 6.1, the performance of models 13-16 reveal important insights into the requirements for a successful predictive model of daily case manager workload. The explanatory power of

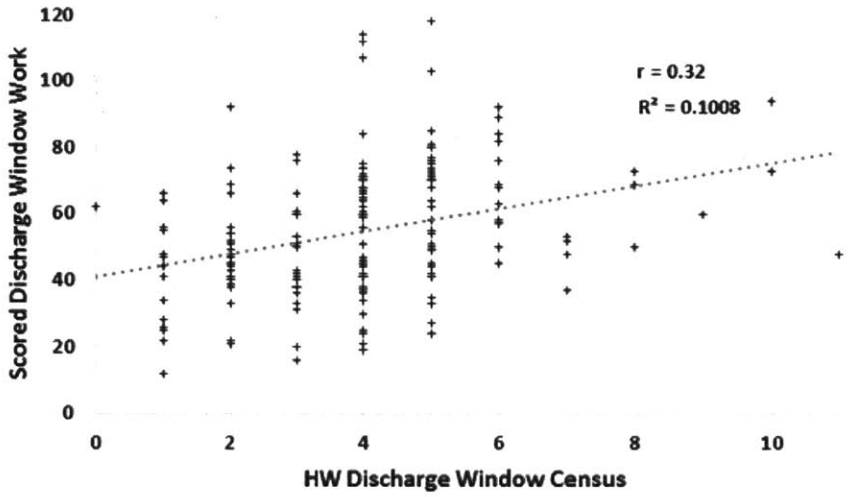
model 13, looking at the daily census count (number of cases, both active and inactive) is virtually non-existent. Of course, this result is expected and it, in large measure, drove our efforts during this work. That the census is a poor predictor of workload is already acknowledged, which is why CMs have varying number of beds they are responsible for based on patient population characteristics. This was discussed in Chapter 2 and generally accepted benchmarks are used to determine how many beds a case manager should be responsible for. Whether these benchmarks are applicable to the CM environment at MGH is an open question, one considered, in part, in the next chapter.

The insights gleaned from the performance of models 13 and 14 are significant in themselves and as guideposts for how the model in section 6.2 must be improved. Recalling Chapter 4, high aggregate workload patients (HW) were defined as the top decile of patients by work scores. This 10% of patients requires 40% of the total work scored over the time periods examined. However, even knowing with 100% accuracy how many HW patients a CM is responsible for on any given day does little to facilitate prediction of daily workload. As model 14 shows, even the number of HW patients in the discharge window does not correlate strongly with daily workload. Finally, though not shown in the table, knowing beforehand which HW cases will be active on a given day offers minimal improvement over a model that only uses the number of active cases as a predictor, making no distinction between HW and not-HW cases. The scatterplots below, showing Pearson's correlation coefficient, r, and R^2 for a one model predictor of daily workload indicates the inadequacy of simply knowing the HW case count – total, discharge window, and active, respectively – for predicting daily CM workload.

Admittedly, expecting a one predictor model to have substantial explanatory power for CM workload is unrealistic, but the more salient point is a model (not shown) with the number of active cases as the predictor has an R^2 of 0.66 in explaining the daily CM workload; a model that segments the number of active cases into HW active cases and other active cases has an R^2 of 0.69 with minimal improvement in RMSE. These results suggest that the timing of work (active cases) and general nature of the cases (admit window or discharge window) offers far more power for predicting workload than identifying HW cases to use as a predictor. At a more basic level, because of low correlation throughout the range of observed values, a high number of HW cases cannot be used reliably as a signal to shift resources from one CM/floor to another even assuming the HW cases could be distinguished from not-HW cases with 100% accuracy. This result, unanticipated beforehand, would have directed our work along a different trajectory. Still, the core of the current predictive model and underlying framework is focused on allowing prediction of the timing of cases and general nature as an implicit step in predicting daily workload.

There are many reasons why the count of HW patients is only weakly correlated with daily workload. It should be obvious that, on average, 60% of the daily workload is from non-HW patients. Yet this does not explain why a large number of HW cases, meaning a number well above, for example, the average or median, is not more strongly correlated with daily workload. The answer lies in the nature of the HW classification; the classification is based on aggregate work completed over the course of a patient's LOS and does not consider the timing of this work. For a different perspective, the work score allows other metrics to be calculated such as high peak workload, high discharge intensity workload, or high aggregate intensity workload. Peak workload would be the highest daily score of workload for a patient during the course of a patient's LOS. Discharge intensity would be the total work completed during the discharge window divided by the length of this window and intensity would be total work completed divided by the LOS. The top decile by one measure may not be the top decile by another measure as shown in Table 6.2. For example of the 123 HW





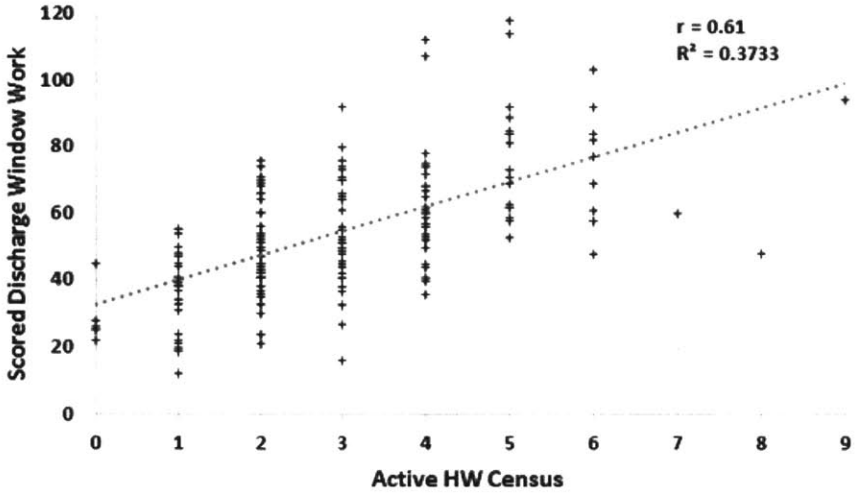


Figure 6-4: The inadequacy of high workload census count for predicting daily workload

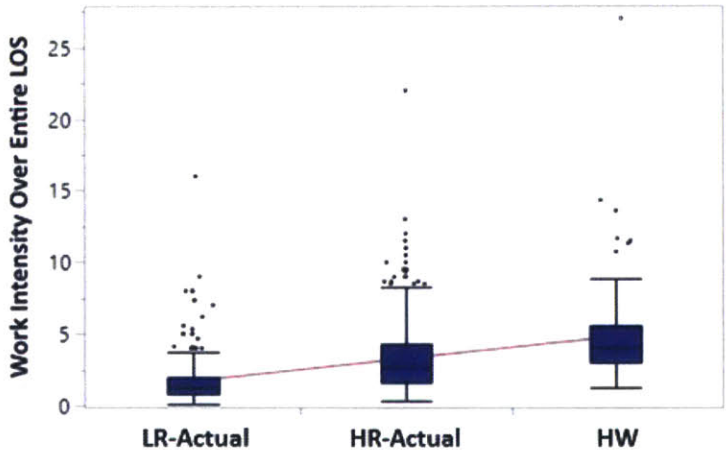
Table 6.2: Shifting membership of top decile groups

	White 8	White 9
HW	123	40
HW + HP	81	27
HW + HDI	36	12
HW + HI	31	13
HW + HP + HDI + HI	23	8

HW *= Top decile aggregate work*
HP *= Top decile peak work*
HDI *= Top decile discharge window work intensity*
HI *= Top decile work intensity over LOS*

patients on White 8, only 31 would be in the top decile of high intensity patients.

Table 6.2 should not be interpreted as meaning HW simply measures an LOS effect. HW patients do tend to have a longer LOS, but as shown in Figure 6-5, the intensity for high workload patients is also statistically higher than not-HW patients. In this figure HR-Actual refers to a patient, exclusive of HW patients, that met high-risk screening criteria and could not, during the initial assessment, be ruled out for needing case manager intervention. LR-Actual patients refer to those patients either not meeting HR criteria, or patients meeting HR criteria that were determined to not require case manager intervention. These categories, in the context of the predictive power of the current high-risk assessment, are considered further in the next subsection, but, in general, LR-Actual patients are low aggregate workload patients and HR-Actual patients are medium aggregate workload patients. The cases considered excluded zero workload cases, patients that had not discharged from White 8 as of 30 June 2015, and patients with a first inpatient department or discharge department different from White 8.



	LR-Actual	*HR-Actual*	*HW*
n	372	425	123
Minimum	0.17	0.29	1.32
10th percentile	0.58	1.00	2.11
25th percentile	0.76	1.58	3.00
Median	1.33	2.67	4.11
75th percentile	2.00	4.29	5.57
90th percentile	3.50	6.50	8.50
Maximum	16.00	22.00	27.00
Mean	1.73	3.32	4.84
SD	1.55	2.46	3.24

Figure 6-5: Work intensity of different patient populations

Table 6.3: Prevalence of a component or subcomponent of work as a plurality of total daily work (holidays and weekends excluded) White 8,1 October 2014 – 30 June 2015

	Days	*Percent of total days*
Admit	**16**	**9%**
Discharge	**171**	**91%**
Admit	56	30%
High workload	43	23%
High discharge intensity	11	6%
High peak	27	14%
Discharge window active yesterday	25	13%
First day of discharge	35	19%
No designation	15	8%

As a group HW patients have higher work intensity, higher discharge intensity, and higher peak workloads than other patient groups, but not necessarily the highest values for these measures on an individual basis. Again, as discussed in Chapter 5 most HW patients require the same types of work, in similar amounts per episode, but repeated multiple times. The request-response cycle, alluded to above, results in a soft cap on the amount of daily work that can be completed for each case so that, on a daily basis, active not-HW cases can impose a similar load as active HW cases. This is not true for idiosyncratic HW cases that make up a fraction of all HW cases.

Finally, on any given day the type of work that predominates may vary. Table 6.3 illustrates this. In this table admit window work and discharge window work are mutually exclusive. The other entries are for components of discharge window work and are not necessarily mutually exclusiveAs the italicized entries are not mutually exclusive, the sum of these entries can exceed 100%; for example, if a patient is in the top decile of peak and total workload, the work performed for that patient would be counted for both the High peak and High workload categories.. In the bottom part of the table admit work is compared to each component of discharge window work, not the sum of discharge window work. The percentage is the number of days in which each component or subcomponent of work was the largest contributor to total daily workload out of 187 non-weekend/holidays considered.

While admit window work is rarely the plurality of work compared to total discharge window work, when compared against subcomponents of discharge window work it is often predominant. The high peak, high workload, and high discharge intensity refer to discharge work completed for the top decile of patients by these measures. "No designation" refers to discharge window work completed for patients that are not in the top decile by any of these measures. The discharge window active yesterday component and its significance is considered in the next section. Again the percent figure given is not for the total amount of work for each component/subcomponent completed over the total number of days considered; it is the percentage of days for which that component/subcomponent dominates, a relevant consideration for predicting daily workload.

6.1.3 Examining potentially exploitable patterns for predicting daily workload

As one may already appreciate, predicting the daily workload of case managers is no easy task. The immediate standard against which the model in the next section was compared was the RMSE and coefficient of determination possible when the number of active admit cases and active discharge cases were used as explanatory variables for total work scored on a given day, 12.44 and 0.65 respectively. Even these modest standards proved elusive, although casting the problem in terms of classification was more promising given the current state of the model. The underlying phased framework for the model is shown in Figure 6-6, with possible inter-day transitions illustrated. Black lines indicate either admissions or movement between phases. Red lines represent the discharge of a patient, with discharges increasing the number of vacant beds on the floor and admissions decreasing this number. In general, a case, from an inter-day perspective, can enter at any phase of discharge planning. Intraday, a patient can pass through multiple phases, but our predictive model assumes counting the number of cases in each state at 04:00 for predictive purposes so intraday transitions are not explicitly captured. Based on the reference work modes in Chapter 5, not all patients progress through all phases (e.g. reference mode 1 patients remain in the unassessed state throughout their LOS, reference mode 2 patients do not progress past pre-discharge, etc.).

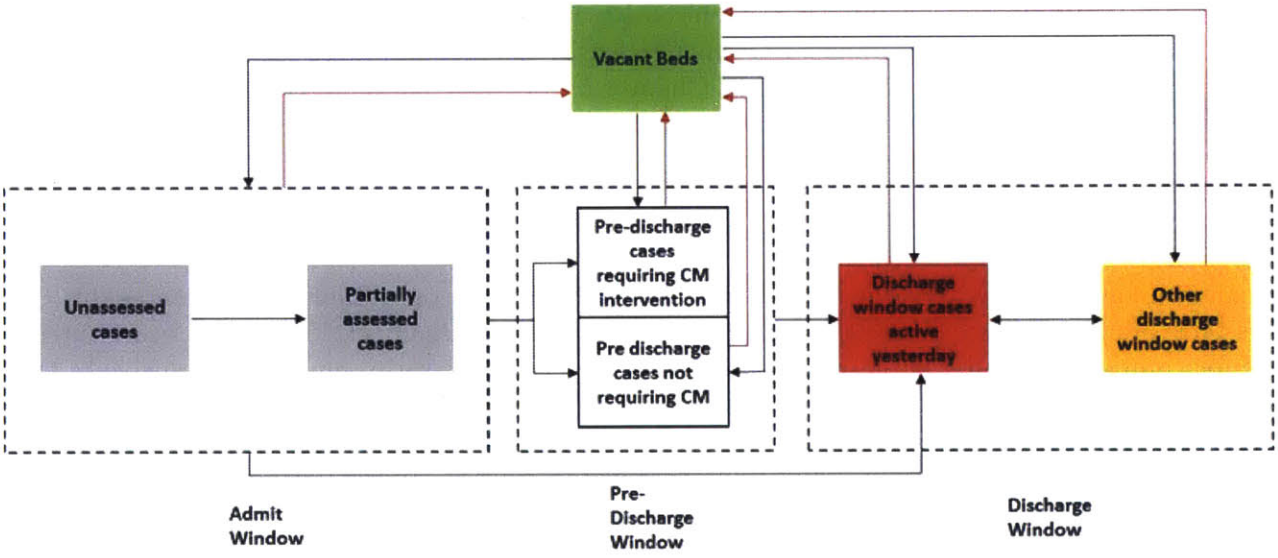


Figure 6-6: Phased structure underlying preliminary predictive modeling

The two potentially exploitable patterns center on cases in the pre-discharge window, distinguishable based on the current high-risk assessment, and cases in the discharge window active the previous day group. By convention, a case in the discharge window on a Monday is counted as active "yesterday" if it was active, in the discharge window, on Friday or the weekend. One of the patterns is, in fact, exploitable – the tendency for cases active the previous day to also be active on the current day. Distinguishing patients in the pre-discharge window based on the HRIA did not prove to be productive for modeling, but this facet warrants a discussion because it touches on the ways that a regression model can be improved by further segmenting the patient population within each phase. In fact, the current preferred model goes beyond the preliminary segmentation of Figure 6-6 by segmenting the discharge window population further, introducing an FDD[3] (patients entering

[3]The models using the FDD count are the exception to the footnote at the opening of this chapter.

Table 6.4: Percentage of total work completed in each phase

	Admit / Pre-Discharge	Discharge window	First / only active discharge day	First active discharge day / multiple days	Discharge window active yesterday	Discharge window not active yesterday
White 8	30.2%	69.8%	11.7%	15.4%	29.0%	13.7%
White 9	32.1%	67.9%	10.8%	15.8%	27.7%	13.7%

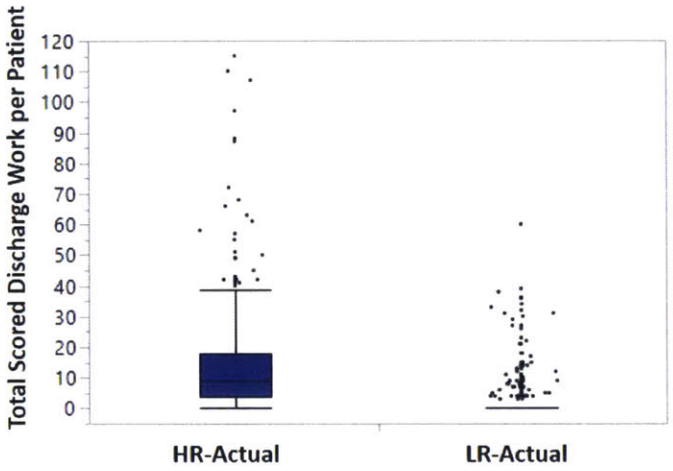
their first active day in the discharge window) count and a "Discharge Window not Active Yesterday" count (DWNAY). The FDD count would have to come from a prediction at the end of the previous business day (before 04:00 on the current day) of which patients will be ready for active discharge planning the following day. The DWNAY count is known with certainty. The FDD count would not be known with 100% accuracy; in fact, it is impossible to say with what accuracy it would be known. The key to obviating the need for a "predicted predictor" like FDD is greater segmentation of the pre-discharge window and even more refined segmentation of the discharge window. This is considered in the next section.

Table 6.4 shows the breakdown of work for patients on White 8 and White 9 by each phased census segment when FDD and DNWAY are included. These amounts differ from Chapter 5 because all patients are considered for the time periods examined, not just patients with White 8 / White 9 as the first inpatient department and discharge department. Note that only a very small amount of pre-discharge work was scored (<1.1% of total) and is included with admit work. The similarity in percentages between White 8 and White 9 is striking.

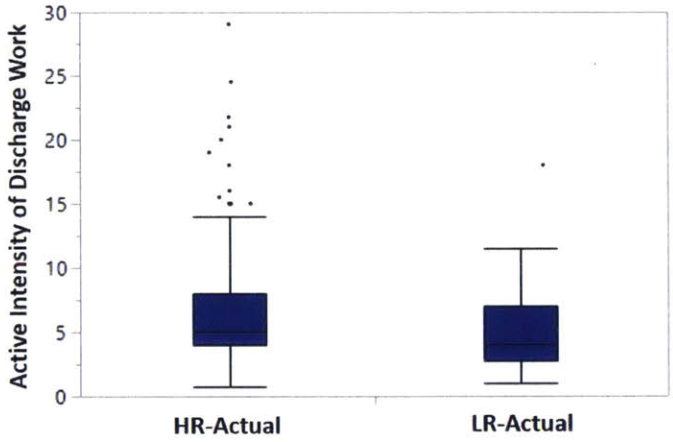
Considering first the segmentation of the pre-discharge window patients, in theory it is possible to segment HR-Actual patients and LR-Actual patients. Here the Actual suffix is used to refer to a patient designation after the CM has completed the HR screen and/or the initial assessment. The LR-Actual patients either fail to meet high-risk criteria or, during the initial assessment the CM determines the patient does not require CM intervention. HR-Actual patients meet high-risk criteria and may need CM intervention. When we overlay the HW classification we developed for the top decile of aggregate workload patients with the results of the HR screen and/or initial assessment, the results are worth noting, as shown in Table 6.5 for White 8 and White 9. In these tables only the patients with White 8 (White 9) as the sole inpatient department are considered.

The most relevant points in the figure are the negative predictive value of the HR screen and the low precision. That is, in terms of high aggregate workload (top decile of patients by work score) the HR screen has a large false positive rate and a very low false negative rate. An even more potentially useful result dispenses with the HW designation and considers the aggregate work required, and other derived metrics, for HR-Actual and LR-Actual patients. This is shown in Figure 6-7 for White 8 patients. Note that the measures refer to discharge window work, hence the presence of zero values.

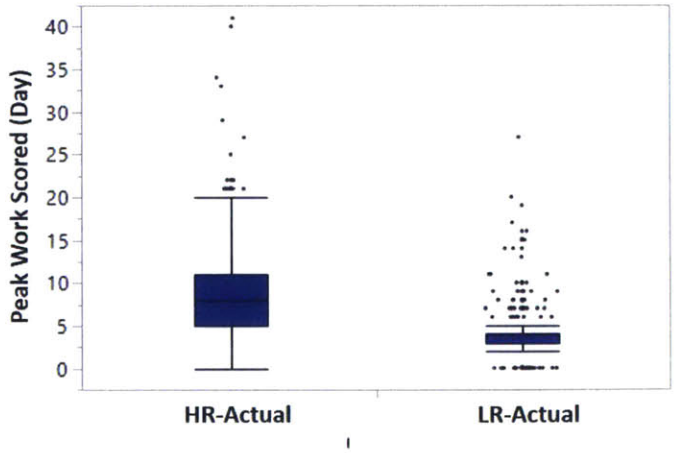
By any metric the current HR screen, combined with the ability of the White 8 case manager to accurately identify patients that do not require case manager intervention, is a powerful distinguisher of different patient populations on the whole. This is potentially useful because many LR-Actual patients do not progress past the pre-discharge window. Our reasoning for segmenting the pre-discharge window as such was that a count of patients in the pre-discharge window, segmented between LR-Actual patients and HR-Actual patients should positively correlate with the number



	HR-Actual	LR-Actual
n	512	410
Minimum	0.00	0.00
10th percentile	0.00	0.00
25th percentile	4.00	0.00
Median	9.00	0.00
75th percentile	18.00	0.00
90th percentile	33.00	9.00
Maximum	115.00	60.00
Mean	13.86	2.87
SD	15.93	7.38



	HR-Actual	LR-Actual
n	426	99
Minimum	0.78	1.00
10th percentile	2.03	2.00
25th percentile	4.00	2.75
Median	5.00	4.00
75th percentile	8.00	7.00
90th percentile	10.77	9.00
Maximum	29.00	18.00
Mean	6.04	4.88
SD	3.76	2.86



	HR-Actual	LR-Actual
n	512	410
Minimum	0	0
10th percentile	4	2
25th percentile	5	3
Median	8	3
75th percentile	11	4
90th percentile	15	8
Maximum	41	27
Mean	8.81	4.04
SD	5.11	3.21

Figure 6-7: Measures of work, work intensity, and peak work for White 8 LR-Actual and HR-Actual patients

Table 6.5: Ability of current high-risk screen and initial assessment to distinguish high aggregate workload patients

White 8

		HR Actual	
		Y	N
High Workload	Y	82	12
	N	513	410

Performance Summary	
Precision	0.14
Recall	0.87
Neg predicitve value	0.97
F-score	0.24
Balanced Accuracy	0.66
G-mean	0.62

White 9

		HR Actual	
		Y	N
High Workload	Y	28	2
	N	107	161

Performance Summary	
Precision	0.21
Recall	0.93
Neg predicitve value	0.99
F-score	0.34
Balanced Accuracy	0.77
G-mean	0.75

of FDD patients for a given day; rather than making a direct prediction of FDD, the prediction was implicit in the model. Unfortunately, Pearson's r for the correlation between the HR-Actual pre-discharge window population and FDD is only 0.454 for White 8; the correlation with non-FDD discharge work is actually negative. Again, the timing of the transition from the pre-discharge window to the FDD is critical to capture given the percent of total work scored on the FDD or the first and only day of active discharge in Table 6.4, an amount in excess of 25%. From Chapter 5 we know that the pre-discharge period is relatively long and latent compared to the admit window and discharge window. Without a finer segmentation the count of pre-discharge patients has limited value for a predictive model. Ideas about how to segment the pre-discharge population in a more granular way are included in the next section. However, only a small amount of work has been done, to this point, on a refined census segmentation scheme; this is why the FDD variable was included as a predictor at this stage of model development.

In contrast to the pre-discharge window segmentation, segmenting the discharge window to include DWAY and DWNAY has useful predictive power, particularly the DWAY category where 29% of the total work scored occurs. On White 8, Pearson's r between the DWAY count and the number of active discharge cases for a day, other than FDD cases is 0.754, with a similar correlation, 0.712, between discharge work performed for active cases other than FDD cases. This type of correlation is highly exploitable for modeling and exists on both White 8 and White 9. The number of contiguous days on White 8 and White 9 is shown in Table 6.6; it is no coincidence that the percentage is close to the Pearson r value for White 8. The tendency of cases active the previous day to be active the current day is a consequence of the request-response cycle mentioned previously. Even with this correlation, further segmentation of the discharge window, in general, and the DWAY population, in particular, would greatly increase the power of the model, as would more explicit modeling of the request-response cycle.

Table 6.6: Frequency of contiguous active discharge days for White 8 and White 9 for patients with varying numbers of total active discharge days

		White 8		White 9	
		n	% contiguous	n	% contiguous
Active days	2	182	69%	40	75.0%
	3	150	69%	54	74.2%
	4	162	68%	54	73.1%
	5	92	73%	40	65.0%
	6+	246	64%	53	60.3%
	Total	832	68%	241	69.5%

6.2 Results of predictive modeling based on linear regression

The results obtained for multiple predictive models, distinguished by the count variables used for prediction, are provided in Table 6.7. These results are based on a model developed with all of the non-weekend/holiday data for White 8. Validation set results, both with repeated 60/40 training/validation trials for White 8 and with White 8 as a training set and White 9 as the validation set are considered below. Only a few of the models in Table 6.7 warrant discussion in terms of performance. A description of the candidate predictor variable codes and response variables is provided in Tables 6.8 and 6.9, respectively. The results are evidence of support that the underpinnings of the model, segmenting the patient population by discharge planning phase to get basic count predictor variables, are solid and worthy of further development. The model performs even better at predicting the active daily census, as number of active admit cases and active discharge case, in lieu of workload. The concept of the active census is one that may hold great promise for future modeling. Finally, the shortcomings of the model naturally suggest improvements.

First, consider model 1, only aimed at predicting admit work. The performance of this model is comparable to model 10 in Table 6.1. Model 10 uses the number of active admit cases to predict admit work with an R^2 and RMSE of 0.78 and 5.29. Model 1 in Table 6.14 uses a 04:00 count of unassessed patients, partially assessed patients, pre-discharge patients, and vacant beds to yield a model with R^2 and RMSE of 0.68 and 6.45. All of the coefficients are significant and, as importantly for extending the predictive model to floors with a different average census, the intercept is not significant. The magnitude of the intercept, relative to the coefficients is, however, potentially problematic for wider application of the model as is.

The relatively good performance of this basic model based on counts illustrates a point obliquely broached earlier. The timing of admit window work should, ultimately, be easier to pinpoint than discharge window work. This is because initial admit window work typically, but not always, occurs within the first 24 business hours of admission. Because this is not always the case the count of unassessed patients, rather than the count of patients admitted (adjusted admit time) within the last 24 hours is used. In contrast the magnitude of admit window work, dependent as it is on the direction of the initial CM encounter with the patient, should ultimately prove to be more difficult to predict, without some type of automated pre-screening, than discharge window work. Admit work usually occurs with limited, or potentially erroneous/incomplete, information about the patient. This limited information is the reason for admit work, be this work a chart review or

Table 6.7: Performance measures of OLS predictive models using White 8 data

		Candidate Explanatory Variables													
Model	Intercept	VAC	UA	PA	PD	PDLR	PDHR	DWAY	DWNAY	FDD	Response Variable	R^2	Adjusted R^2	RMSE	Mean Response
1	-4.22 *3.77*	0.94* *0.37*	3.79* *0.23*	3.49* *0.82*	0.43* *0.21*						Admit Work	0.68	0.67	6.45	*25.03*
2	29.07* *4.72*						1.28* *0.53*	6.30* *1.28*			Discharge Work	0.28	0.27	16.78	*54.41*
3	49.24* *4.00*		2.66* *0.50*					5.74* *0.78*			Total Work	0.28	0.27	17.81	*79.44*
4	15.34* *4.03*							6.69* *0.64*	1.47* *0.52*	5.50* *0.63*	Discharge Work	0.48	0.47	14.34	*54.41*
5	29.8* *4.3*		3.27* *0.44*					6.35* *0.68*		5.27* *0.68*	Total Work	0.45	0.45	15.52	*79.44*
6	2.09* *0.52*		0.75* *0.04*					0.97* *0.07*	0.19* *0.06*		Active Cases not FDD	0.66	0.66	1.62	*9.70*
7	1.11* *0.25*		0.69* *0.03*	0.46* *0.17*							Active Admit Cases	0.65	0.66	1.37	*5.19*
8	1.15* *0.28*							0.90* *0.05*	0.23* *0.04*		Active Discharge Cases not FDD	0.60	0.59	1.24	*4.51*
9	2.32* *0.18*							0.83* *0.06*			Active Discharge Cases not FDD	0.54	0.54	1.32	*4.51*
10	0.25 *0.45*				0.22* *0.04*						FDD	0.15	0.15	1.57	*2.72*
11	0.09 *0.44*					0.11* *0.05*	0.31* *0.05*				FDD	0.19	0.19	1.52	*2.72*

** Significant at $\alpha < 0.05$*

*** Significant at $\alpha < 0.10$*

Table 6.8: Description of candidate predictor variables

Variable Code	Description*
VAC	Vacant beds on floor
UA	Unassessed patients on floor
PA	Partially assessed patients on floor
PD	Pre-discharge patients (assessed but active discharge planning has not started
PDLR	Pre-discharge patients who were assessed as low-risk actual
PDHR	Pre-discharge patients not assessed as low-risk actual
DWAY	Discharge window patients with active discharge planning previous day
DWNAY	Discharge window patients without active discharge planning previous day
FDD	Patients who will enter first day of active discharge planning

**All counts are at 0400 of the same day for which work is scored*

an initial meeting with the patient. Of course, currently the magnitude of discharge work is difficult to predict but there are many opportunities to incorporate new information about the state of the patient that can enhance prediction. The basis for these opportunities is considered below.

Model 5 is the current best model for predicting total work with an R^2 of 0.45, an RMSE of 15.52, and a mean response of 79.44. It bears noting that including a dummy indicator variable for Friday pushes the coefficient of determination for the model above the oft-cited, but ultimately irrelevant, 0.50 threshold for predictive models in the social sciences. The relative RMSE of 20% may preclude the use of the current model to make accurate operationally useful predictions of case manager workload but, as shown in the next section, the model does allow "reasonably" accurate classification of high, medium, and low workload days with an intricate method of classifier training. This method may not be extensible to other floors, but even the current regression results and classifier results, based on simple count variables of census by phase, suggest the core concepts of the model are sound. The difference in relative performance metrics between models 3 and 5 again emphasizes the importance of predicting the timing of work; i.e. predicting whether a case will be active on a given day. Model 5 incorporates the FDD count variable, implicitly assuming 100% accuracy, while model 3 (or model 2) is devoid of a mechanism to accurately capture the transition of patients from the pre-discharge window to active discharge planning. Note that the DWAY variable is significant for all models predicting either total work or discharge work.

Models 6, 7, and 8, predicting active cases not FDD, active admit cases, and active discharge case not FDD, respectively, are particularly provocative. Rather than casting the problem as predicting workload, conceptualizing the problem as one of predicting the active census may result in a more tractable problem. As has been demonstrated, in a manner we feel is conclusive, predicting the timing (active case or not active) and general nature of work (admit window work or discharge window work), is a necessary, even if only implicitly, condition for predicting workload with the degree of certainty needed to operationalize a predictive model. Model 6, for example, allows prediction of active cases not FDD with a RMSE of 1.62 cases. If a prediction of FDD cases is incorporated to make a prediction of all active cases the RMSE drops further by an amount correlated with the accuracy of the FDD prediction.

Table 6.9: Description of response variables

Response Variable	Description
Admit Work	Scored admit work for day (admit window*)
Discharge Work	Scored discharge work for day (discharge window)
Total Work	Total work scored for day
Active Cases not FDD	Number of active cases for day except first day of discharge cases
Active Admit Cases	Number of active admit cases for day
First Day of Discharge (FDD)	Case entering first active discharge day
Active Discharge Cases not FDD	Number of active discharge cases for day except first day of discharge cases

** Includes <1.1% pre-discharge work*

Table 6.10: Performance of OLS models on various validation data sets

		White 8 *All non-weekend/holiday data*			***White 8***** *Repeated trials 60/40 Train/Validation*				***White 9*** *Validation*			
Model	**Response Variable**	R^2	**RMSE**	**Mean Response**	R^2	**RMSE**	**Mean Response**	**R/R RMSE***	R^2	**RMSE**	**Mean Response**	**R/R RMSE***
1	Admit Work	0.68	6.45	25.03	0.68	6.52	25.10	1.01	0.58	12.00	22.73	2.05
2	Discharge Work	0.28	16.78	54.41	0.28	16.14	53.92	0.97	0.25	16.62	44.23	1.22
3	Total Work	0.28	17.81	79.44	0.28	17.13	79.03	0.97	0.22	22.27	66.97	1.48
4	Discharge Work	0.48	14.34	54.41	0.48	13.44	53.92	0.95	0.45	13.20	44.23	1.13
5	Total Work	0.45	15.52	79.44	0.45	14.61	78.83	0.95	0.44	17.97	66.97	1.37
6	Active Cases not FDD	0.66	1.62	9.70	0.66	1.56	9.80	0.95	0.59	2.38	8.92	1.60
7	Active Admit Cases	0.65	1.37	5.19	0.65	1.38	5.21	1.00	0.57	2.29	4.78	1.81
8	Active Discharge Cases not FDD	0.60	1.24	4.51	0.6	1.23	4.58	0.98	0.55	1.40	4.14	1.23

** This refers to the relative RMSE of a validation set* ***relative*** *to the relative RMSE of the White 8 models developed with all of the non-weekend/holiday data*

*** Results shown are the average values obtained for the validation set, 10 trials*

6.2.1 Validation of the model for White 8 and testing extensibility to other floors

The middle portion of Table 6.10 shows the validation results (averaged) of repeated trials where the White 8 data was randomly split into a training set with 60% of the data and a validation set of 40% of the data. The deficits of the current model were discussed earlier, but the model performs equally as well, identically in fact, on White 8 validation sets.

The rightmost columns of the figure show the performance of a model trained with the entire non-weekend holiday White 8 data set and tested on the entire White 9 data set. Immediately we can see something is amiss. This is demonstrated most clearly in the rightmost column highlighted in yellow. Here we introduce the R/R RMSE measure. Given the different mean responses of White 8 and White 9, this is an appropriate measure to consider as it is the ratio of the White 9 relative RMSE and the White 8 relative RMSE. Comparable performance would result in a value close to 1. The decrease in performance for model 2, model 4, and model 8 is partly attributable to the slightly varying patient population on White 9 compared to White 8 over the time periods examined, as discussed in Chapter 5. This influence is slight because the patient population variation is most

pronounced among patients in the pre-discharge window, a count variable not included in the models. Note that the models where performance tends to agree concern discharge work or discharge cases.

In contrast, the performance of models predicting total work, admit work, total cases, or active admit cases is drastically reduced when tested on White 9. This reduction in performance traces to different work patterns on White 9, as well as a different relative prevalence of work reference modes, also discussed in Chapter 5. These models all make use of the VAC, UA, or PA variables. The admit work is delayed, on average, for White 9 compared to White 8. Thus, the value of the UA coefficient and the VAC coefficient would be decreased in a model tailored for/trained on White 9 data. Similarly, the PA variable, because of the higher relative occurrence of reference mode 7, 8, and 9 cases on White 9 would assume greater prominence in a White 9-trained model. This is a dramatic demonstration that, even if the improvements to modeling outlined in the next subsection are completed, a level of standardization across floors is requisite for a widely applicable predictive model. The alternative is to train models for a given floor only using data from that floor. This would eliminate the effect of different work patterns and role boundaries that may exist between floors.

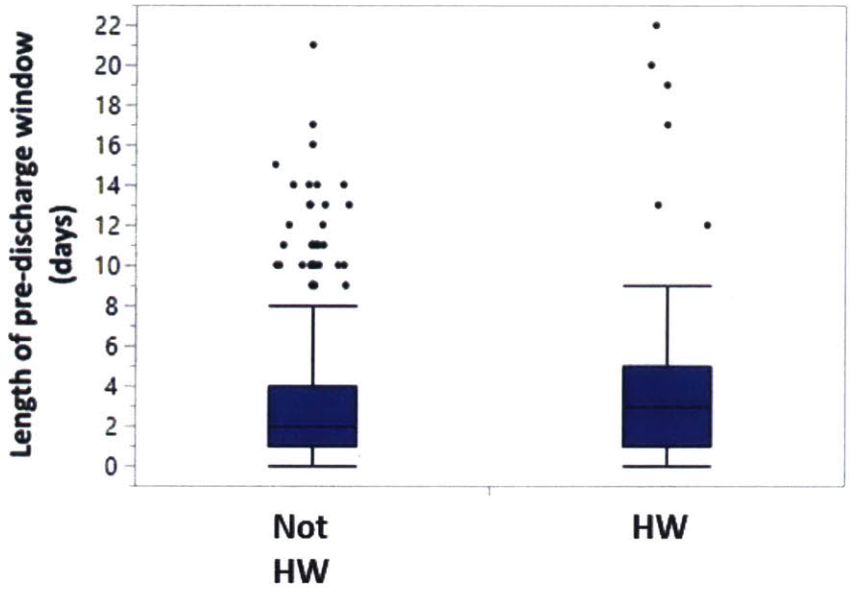
However, without standardization, model drift is virtually guaranteed. On the one hand, processes evolve at a different rate and in different directions depending on the floor. Yet, the real reason standardization is beneficial has to do with all of the improvement initiatives at MGH[4]. These initiatives could easily change the foundation upon which a model is built. If every floor has a slightly different model then changes, internal or external to MGH (e.g., an external change could be healthcare regulation or Medicare changes) would mean all of the models have to be changed. With standardization, at least between similar floors, this situation is mitigated, to a degree.

6.2.2 Improving regression model predictive performance

There are opportunities to enhance the ability of a model to predict both the timing and the magnitude of work by incorporating new, relevant information about the patient as the patient's LOS unfolds. In fact, this should be a much more productive approach, from a predictive standpoint, than trying to make a single prediction of aggregate work at the outset of a patient's stay. For example, consider the pre-discharge window, largely devoid of CM activity for a case as per our definition. The length of this pre-discharge window for HW and not-HW patients is shown in Figure 6-8.

The median (or average) pre-discharge window for both patient groups is non-trivial. Currently, from a predictive modeling standpoint, the pre-discharge window is essentially undifferentiated. However, at a minimum there should be an early pre-discharge window and a late pre-discharge window. Rather than looking at CM activities to distinguish these periods, preliminary investigation indicates that the occurrence of specific consults, particularly physical therapy, could be used to separate a patient in a hypothetical early pre-discharge window and late pre-discharge window. These consults are potentially leading indicators that the phase of active discharge planning (case entering the discharge window) is approaching. This would allow a two-axis segmentation in the pre-discharge window along an early/late axis and a HR-Actual/LR-Actual axis. What is more, the

[4]This is not idle speculation; the change to a new model of case management introduced subtle but discernible discontinuities into the CM note text data at the heart of our analysis. This limited the amount of historical data for each floor available for use in our analysis.



	HW n=117	Not HW n=957
Minimum	0	0
25th percentile	1	1
Median	3	2
75th percentile	5	4
90th percentile	7	6
Maximum	22	21
Mean	3.72	2.64
SD	3.87	2.74

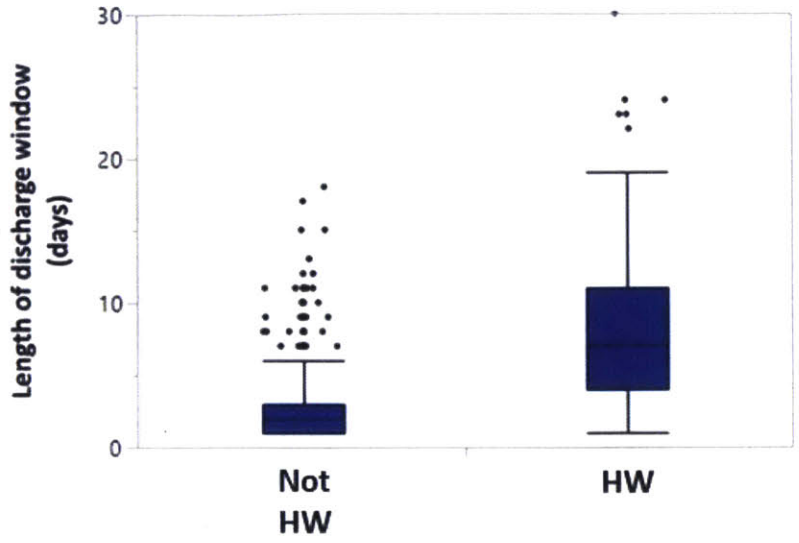
Figure 6-8: Length of pre-discharge window for HW and not-HW patients, White 8, 1 October 2014 – 30 June 2015

classification of patients need not remain static as new information about the patient's psychosocial and insurance profile becomes available. A great deal of effort was spent in trying to develop a method to predict a HW patient in the early phase of admission. Much more beneficial for predictive modeling of daily workload is incorporating information as it becomes available. Appendix A briefly summarizes earlier efforts at predicting high workload patients early in admission. These same concepts, because of relatively inactive periods during a patient's LOS, from a CM perspective, can be applied throughout the LOS to refine the prediction of work timing and magnitude.

As one final illustration, consider an oft-cited obstacle to predicting the amount of work associated with a patient, a frequently shifting discharge disposition. This change can be because a patient's needs change or because a patient's psychosocial characteristics or insurance profile means patient needs must be met in a manner other than that preferred by the patient and initially pursued by the CM. Figure 6-10 illustrates this fact. This figure is made difficult to read because of the presence of so many outliers. In fact, these outliers are the key feature of the figure in that there does not seem to be any well-defined characteristic amount of work or number of active discharge days for an ultimate discharge disposition. However, it is our contention that there is a characteristic amount and sequence of work for the **discharge disposition being pursued at the time.**[5]

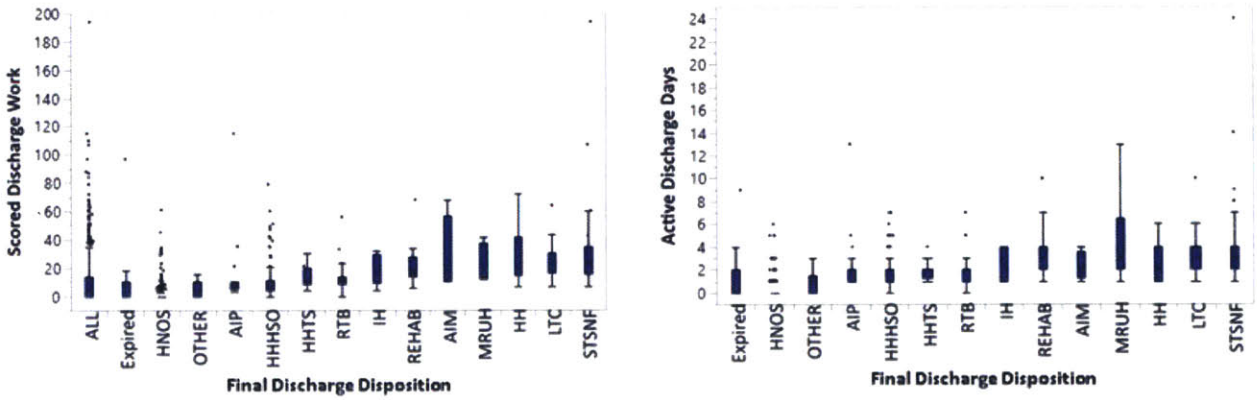
Consider the Home-Services (HHHSO) category. These patients typically require a lower level of CM intervention. This fact, at first glance, does not seem to square with the figure's outliers for this category. The two observations, however, are not contradictory and are easy to reconcile after an exhaustive reading of CM notes. The outliers for this category are typically patients for which

[5]Another way of stating this assertion is that if all patients with the same ultimate discharge disposition started initially along a path to this ultimate discharge disposition (no shift in the pursued discharge plan), the observed variation evident for each category of ultimate discharge disposition would be lower. The characteristic amount of work for the first day of pursuing a type of discharge disposition, for example, would vary in a narrow range.



	HW n=123	Not HW n=524
Minimum	1	1
25th percentile	4	1
Median	7	2
75th percentile	11	3
90th percentile	19	6
Maximum	98	18
Mean	9.76	2.65
SD	11.12	2.53

Figure 6-9: Length of discharge window for HW and not-HW patients, White 8, 1 October 2014 – 30 June 2015



		Discharge Work					Active Discharge Days				
	n	Min	Median	Max	Mean	SD	Min	Median	Max	Mean	SD
All	1077	0.0	4.0	194.0	9.94	0.43					
Expired	17	0.0	0.0	97.0	9.76	3.43	0.0	0.0	9.0	1.24	2.39
Home - No services (HNOS)	472	0.0	0.0	61.0	1.39	0.65	0.0	0.0	6.0	0.21	0.69
Other (OTHER)	13	0.0	3.0	15.0	4.77	3.92	0.0	1.0	3.0	0.92	0.95
Acute inpatient - Psychiatric (AIP)	25	3.0	6.0	115.0	12.24	2.83	1.0	2.0	13.0	2.36	2.43
Home - Services (HHHSO)	273	0.0	7.0	79.0	10.07	0.86	0.0	1.0	7.0	1.50	1.10
Home - Hi-Tech Services (HHTS)	12	4.0	10.0	30.0	12.92	4.08	1.0	2.0	4.0	2.00	0.85
Return to admitting facility (RTB)	45	0.0	10.0	56.0	13.38	2.11	0.0	2.0	7.0	1.84	1.24
Hospice - Inpatient (IH)	6	4.0	17.0	32.0	18.17	5.77	1.0	2.0	4.0	2.33	1.37
Rehab (REHAB)	28	6.0	17.5	68.0	22.75	2.67	1.0	3.0	10.0	3.21	2.10
Acute inpatient - Medical (AIM)	4	10.0	18.0	68.0	28.50	7.07	1.0	2.0	4.0	2.25	1.26
Medical respite unit - homeless (MRUH	13	12.0	19.0	110.0	32.92	3.92	1.0	3.0	13.0	4.38	3.86
Hospice - Home (HH)	18	7.0	19.5	84.0	30.17	3.33	1.0	2.0	6.0	2.56	1.62
Long-term acute care (LTC)	32	7.0	22.0	87.0	25.13	2.50	1.0	3.0	10.0	3.38	1.72
Short-term SNF (STSNF)	124	7.0	22.5	194.0	27.61	1.27	1.0	3.0	24.0	3.32	2.69

Figure 6-10: Total discharge window work and active discharge days associated with ultimate discharge dispositions, White 8 1 October 2014 – 30 June 2015

a different discharge disposition (or dispositions) was initially envisioned but whose post-MGH placement proved so difficult that, eventually, the patient's condition improved to the point that a discharge home with services was possible. In fact, many of the outliers, particularly in the STSNF (short-term skilled nursing facility) category may have had partial discharge plans executed for a number of different discharge dispositions. The point is that, at any point in time the discharge disposition being pursued could be included in a predictive model, as could outcomes, such as multiple denials by post-MGH facilities. Incorporating this outcome-based type of information is the key for near real-time (daily) predictive modeling of CM workload and allows a finer segmentation of the current census by phase and sub-phase.

6.3 Reformulating the daily workload prediction problem as a classification problem

As discussed, predicting an accurate score for the daily workload of a case manager is difficult. There are avenues to improve the current predictive model but, from a near-term pragmatic perspective, attempting to accurately classify a day as high, medium, or low workload may be a better course to pursue. A variety of classifiers were examined (logistic regression, knn, random forests, svm, discriminant analysis, neural nets, etc.) but ultimately a boosted tree-based classifier was chosen. Since daily workload classification attempts were started relatively late in the process the parameters of the boosted tree were minimally tuned.[6] Fifty weak learners were used with SMOTEBoost as the boosting algorithm for a standard CART implementation[44][55]. A terminal node was constrained to have a minimum of 10% of the training set days. The details of the tree are really ancillary to this discussion as an intricate stratified random sampling procedure offered the greatest benefit to classifier performance.

Stratified random sampling went well beyond simple stratification based solely on class. The techniques are admittedly non-standard and some results, particularly with the small White 9 data set, give reason for pause. As such, the techniques used below should be verified on a larger, novel data set. The advantage of the linear regression model is its transparency and relative simplicity. The modifications suggested in the previous section should result in improved performance and the reason for any performance improvement is easily conceptualized. In contrast, there is a certain opacity associated with using techniques like hierarchical clustering, boosted classification trees, synthetic minority class over-sampling technique (SMOTE), and one-side selection in combination.

The sampling procedure began with hierarchical clustering. All variables referenced in Tables 6.1 and 6.7 were used for clustering. Three clusters were formed. The data points (days) in these clusters were then assigned to six sub-clusters by manual clustering based on the dominant work components in the cluster from Table 6.3 among admit, high peak, first day of discharge, and DWAY with the restriction that a minimum number of high workload cases was in each cluster[7].

[6]There is an opportunity to use a wrapper technique for tuning over a range of values for the parameters associated with the approach described in this section[62][53][80]. Essentially this would entail a grid search to determine the optimal sampling method for our training dataset. While simple in concept, wrappers can become complicated when trying to optimize a pair of over- and under-sampling levels.[80][57]

[7]This clustering was entirely manual and based on little more than a belief that the SMOTEBoost algorithm could be applied to a training set of data and result in a reasonably well-performing predicitve classifier. Six sub-clusters were chosen because this was the maximum amount of sub-clusters meeting a minimum cluster size requirement of four high workload days per cluster; four was set as a minimum to allow use of the SMOTEBoost technique on each

The data set was then split into a training and test set by randomly taking half of the high work days from each cluster (or half – 1 in the case of an odd number) for the training set and the remaining cases for the test set. The days with a classification other than high were similarly split 50/50 among the training and validation set to yield a six strata training set and validation set.

The SMOTE technique was then used on the training set to form an equal number of synthetic high workload data points in each strata, assigning composite values of the variables based on the five nearest neighbors, whether high workload or other, of each high workload day. A significant amount of data munging was required to maintain the internal consistency of the training data set. For example, if there were five patients in the discharge window then there could not be six DWAY cases for a synthetic high workload day. The one-sided selection algorithm was then used to remove non-high workload days in the training set based on Tomek links and condensed nearest neighbors[125][77][87]. Finally, the boosted regression tree was built based on the remaining training data using the full set of candidate predictor variables only (Table 6.7). The process was repeated for 11 trials. The SMOTEBoost technique ensured some variation between trials despite the small number of high workload days.

With the reservations about extensibility of the stratified sampling procedure duly noted, the validation results for White 8 are shown in Table 6.11.

Table 6.11: Validation set performance for High Workload Day / Not-High Workload day classifier on White 8

		Predicted Class	
		H	L
Actual Class	H	87	34
	L	32	661

Performance Summary		
Precision		0.73
Recall		0.72
F-score		0.73
Balanced Accuracy		0.84
G-mean		0.83
ROC Curve, AUC (11 Trials)		
	Average	0.87
	Median	0.88
	Range	(0.81, 0.91)

By some metrics the classifier performance can be judged quite favorably. The values are summed over 11 trials. The real context in which to judge performance depends on the operational cost of misclassification. This cost is unlikely to be symmetric. The classifier performance is relatively balanced along the off-diagonal, misclassifying 34 high workload days as low workload and misclassifying 32 low workload days as high workload. The 34 misclassifications are opportunities missed where resources could have been reallocated to make a CM's daily workload more manageable. Conversely, the 32 misclassifications could result in limited reserve capacity being flexed to a floor that does not require additional resources. In general we can trade off an increase in true positives only by accepting a higher number of false positives. The "median" ROC curve and lift charts

cluster in forming an augmented training set.

are shown in Figure 6-11. The lift chart shows that the highest workload days are identified with relative ease.

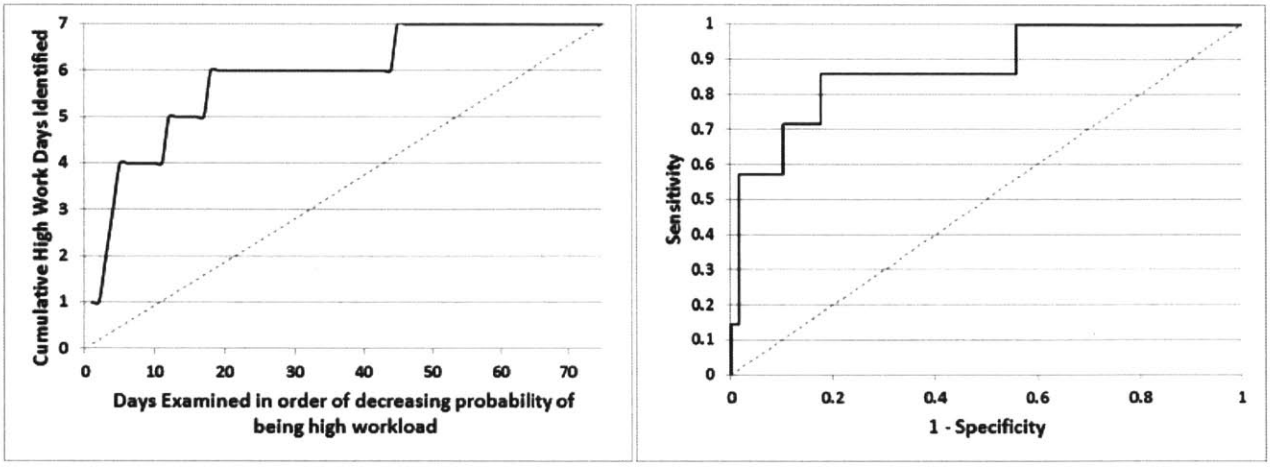


Figure 6-11: Median validation set lift chart and ROC curve for White 8 two-class classifier

The results of training a classifier with a White 9 training set and validating on a White 9 holdout set are shown in Table 6.12, with accompanying median lift charts and ROC curves in Figure 6-13. The same procedure was used as described previously, with 11 trials completed. The number of high workload days on White 9, was smaller so the cutoff score was lowered to 90. Still, the White 9 data set is only 1/3 the size of the White 8 data set (64 days versus 187 days) and the classification, even augmenting the training data set using the techniques described above, is not likely to be indicative of performance on a larger data set; the performance on a larger data set could be better or worse. This classifier did not exhibit the off diagonal balance of the previous classifier, having comparatively more false positives than false negatives. More notable is the instability in the ROC curve between trials. The median and average value was similar to the previous classifier but the range, with a low AUC of 0.59, was much wider. Again, it is difficult to read much into the performance metrics because of the small sample.

Table 6.12: Validation set performance for High Workload Day / Not-High Workload day classifier on White 9

		Predicted Class	
		H	L
Actual Class	H	30	14
	L	29	202

Performance Summary		
Precision		0.51
Recall		0.68
F-score		0.58
Balanced Accuracy		0.78
G-mean		0.77
ROC Curve, AUC (11 Trials)		
	Average	0.81
	Median	0.83
	Range	(0.59, 0.93)

Finally, Table 6-13 provides the classification results for 11 trials of a three class classifier with break points at 60 and 90 to distinguish between the classes. The combined data for White 8 and

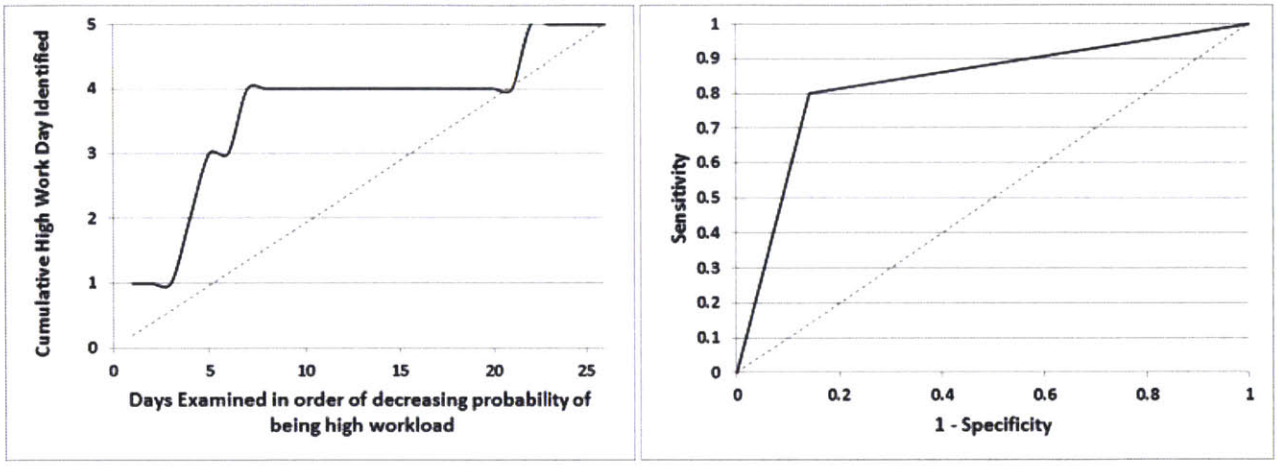


Figure 6-12: Median validation set lift chart and ROC curve for White 9 two-class classifier

Table 6.13: Validation set performance of a 3-class classifier over 11 trials

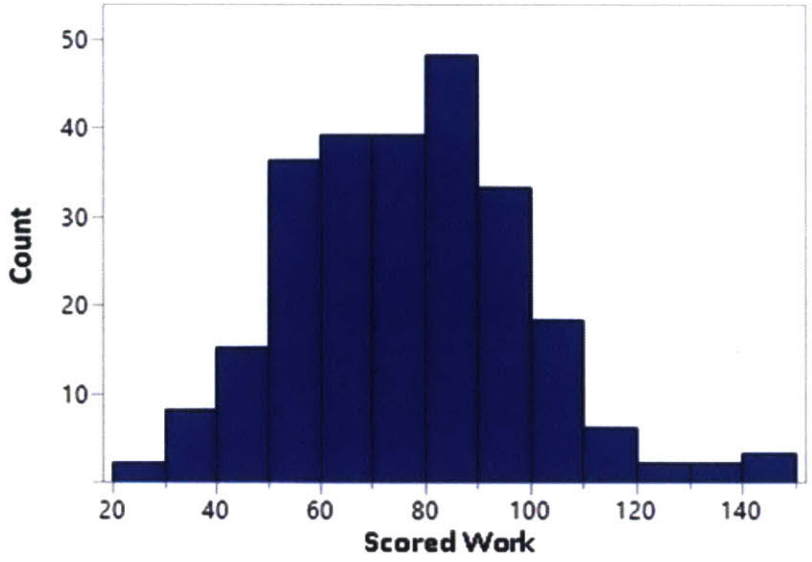
Actual Class	Predicted Class: H	Predicted Class: M	Predicted Class: L
H	107	27	10
M	32	238	60
L	14	51	491

Performance Summary	
Overall Accuracy	81.2%
High Load Day Accuracy	74.3%
Medium Load Day Accuracy	72.1%
Low Load Day Accuracy	88.3%
Two-level miclassification rate	3.4%
High as low misclassification	6.9%
Low as high misclassification	2.5%

White 9 was used in a manner similar to that described at the outset of this section, although three classes necessitated a modification of the hierarchical clustering-based stratified random sampling. The important practical metric for this classifier is the overall two-class misclassification rate and the component low-as-high and high-as-low misclassification. An overall two-class misclassification rate of 3.4% was obtained.

Like the regression based model, the preliminary classification results show promise. However, it is not as clear whether the classification results, dependent as they were on a complicated stratified random sampling procedure to properly train a classifier, will extend to novel data sets. Furthermore, classification requires a better calibration of the work score scale to determine exactly what score constitutes a high workload day. The distribution of weekday scores is shown in Figure 6-13. The same problem exists for the regression-based model but calibration does not have the same immediacy since refinements to the model can be made and evaluated relatively independently of the scale. The simplicity of the regression model is a huge attraction, although the refinements likely needed to improve the model to the degree necessary will complicate the model marginally. It is probable that any improvements will likely translate to faster performance enhancements for the classifier. At any rate, both avenues are open for exploration. The final chapter considers more immediate operational improvements that may be possible, dependent mainly on the retrospective work score for support rather than a predictive modeling capability. Chief among these are the effects of pooling the work of similar floors between the associated case managers and simple procedures that may provide a

basis for altering the baseline staffing among floors with different patient populations.



Stem	Leaf	Count
14	699	3
13	37	2
12	07	2
11	011446	6
10	000134444455668899	18
9	000111122222223333445555556688899	33
8	011111122223333344555555666666777778888888899999	48
7	000111112222233444444455566677888899999	39
6	000111222233344445555666666778888899999	39
5	000111122233344556667777777888899999	36
4	003334456888899	15
3	35778899	8
2	89	2

Figure 6-13: Distribution of daily workload scores on White 8 and White 9

Chapter 7

Recommendations for Future Work and Operationalization Roadmap

A number of recommendations for future work in the context of quantifying CM work, and more accurately predicting the daily workload for inpatient case managers at MGH, have been presented throughout Chapters 4-6. The most important of these recommendations are recounted, briefly, below, and focus primarily on ways to improve the prediction accuracy of the daily workload required for a set of cases.

While the current state of our work does not allow for operational recommendations to be made with the requisite specificity for immediate implementation, it is possible to move beyond, in many respects, vague generalities. This allows us, at the very least, to identify the linkages between the recommendations for future work and operationalization of this work, to provide a roadmap, if you will, for operationalization.

Appendix B outlines possible ways to use both the workload metric and/or our phased framework to develop a CM-specific case mix index (CMI). The appendix provides more information about the case mix index for readers unfamiliar with the term. The concept is relatively straightforward – the CMI assigns a relative value to different patient groups in an effort to determine how hospital resources should be allocated among these groups, as well as how many resources patients in a particular group typically require. As we view case managers and their time as an important hospital resource, we consider adapting the CMI concept to develop benchmark caseloads for different patient populations / hospital floors.

Of course, as we have demonstrated in the preceding chapters, even with better, MGH-specific, benchmark caseloads, sub-optimality in the daily allocation of case management resources is a virtual certainty. That is, while benchmark caseloads can help ensure an equitable distribution of workload over an extended time horizon (i.e., on "average"), on a daily basis the sources of workload variability can have significant effects. In this chapter we introduce a feasible dynamic case assignment scheme, based on pooling to help limit the upside daily workload variability experienced by any case manager within the pool. The attenuation can be achieved by balancing the number of cases in each phase between case managers within a pool.

7.1 The fundamental importance of the work metric and phased framework: Suggested refinements

As stated in the introduction to Chapter 4, the foundation of our work hinges on the work metric developed. This work metric is key to any meaningful current state analysis, as well serving the role of response variable for analyzing CM note text, explanatory modeling, and predictive modeling. Given this importance, it is imperative that further refinement and validation of the metric occur, particularly for making comparisons across floors with different patient populations. The relatively small sample of cases used to validate the work metric is the primary weakness of the work to date. There are several reasons for the small sample, chief among them being the time-consuming nature of reviewing cases. However, involving a larger number of auditors, reviewing a larger number of cases, to arrive at a consensus scoring system would undoubtedly be more efficient than a protracted time-motion study approach.

Rather than classifying cases according to how much work they required (e.g., high, medium, or low), or ordinally ranking a sample of cases by the amount of work required, auditors should explicitly score the text of cases, with the goal of calibrating the scale for the work metric in terms of time required. This could begin at the level of component work events, such as phone calls, referrals, or faxes, but more beneficial, in terms of refining the text-analytical engine that will form the core for automated scoring, would be scoring at the note-level. This facilitates enhancements to automated scoring by allowing easier extraction and incorporation of textual markers associated with a given level of work into the automated scoring procedure. The note type would remain indispensable as a feature of a text vector to be scored, but consensus scoring of text samples at the note level decreases the importance of specifying beforehand the dictionary presented in Appendix C. More accurately, a significant sample of notes scored by a group of case managers provides a more substantial foundation for any semi-supervised learning, via clustering, used to guide development of the dictionaries and sub-dictionaries in Appendix C.

There are other refinements that a more substantial audit of cases and scoring of notes may make attainable. This audit procedure may allow estimation of the effects of having multiple active cases per day so the effects can be incorporated into retrospective scoring. Some of these effects include the time cost of switching from working on one case to working on another. The case managers we observed are adept at switching between cases, but this is not a frictionless process. This could be an important effect to consider as it likely means that the linearity assumed in developing the regression-based models of Chapter 6 holds only over a limited range of active cases. Even more significant is the increased chance of interruption by other members of the care team as the number of active discharge window cases increases. These interruptions were frequently observed, but incorporating them into a case score, particularly an automated case score, is difficult. At the day-level this leads to a downward bias in scoring; at the patient level this similarly leads to a downward bias in the total case score and, more importantly in the context of daily predictive modeling, lower daily peak scores for a case.

Another primary benefit that could come out of a directed, larger-scale review of cases is a refinement of some of the key conventions developed over the course of our work. Chief among these is the basic phased framework used to identify the phase a case is in, the basic phases being admit window, pre-discharge, and discharge window. Within these general phases we further pinpointed the state a case is in, such as UA, PA, or DWAY. Identifying on any given day how many cases were in each

state forms the core of the current predictive model and will form the core of future predictive models, whether regression-based, classification based, or otherwise. This segmentation of cases by phase, in a manner not informed by our early working hypothesis, proved much more beneficial for predictive modeling than, for example, identifying aggregate high-workload cases. The reason for this was clearly demonstrated in preceding chapters – the timing of work (is a case active or not?) and the general nature of the work (which phase is the case in?) goes far in allowing prediction of the daily workload.

A more granular segmentation of cases, up to a point, would allow even better prediction of daily workload, both by improving prediction of the timing of the work and allowing a better prediction of the magnitude of the work. As an example of the former, there may be identifiable phases within the pre-discharge window, with transitions indicated by sources other than CM notes. As mentioned in Chapter 5, it may be possible to distinguish between an early and late pre-discharge window based on the occurrence of certain consults. If distinguishable phases in the pre-discharge window are identified then it may be unnecessary to predict which cases will be FDD cases during the upcoming day (although incorporating a prediction of FDD case count will almost certainly improve a predictive model). Refined segmentation may allow the negative predictive value of the current HRIA to be exploited for modeling purposes. Furthermore, it is likely a more refined segmentation would allow better incorporation of the request-response cycle into a predictive model of daily workload, a cycle which the DWAY state does an admirable, though incomplete job of capturing.

As an example of segmentation that may allow better prediction of the magnitude of work associated with a case on a given day, a count of patients by predicted discharge disposition could be beneficial. More beneficial would be capturing the relevant outcomes of the previous day's work on a case. For example, was a patient/case in the discharge window denied by a facility yesterday? In the same vein, are there reasons for denial that are important indicators for the work that will be completed for a case the coming day? Here input from case managers during a review of cases would be invaluable. Incorporating new information about the state of a case is more useful from a prediction standpoint than an accurate prediction of the aggregate amount of work a case will take from admission to discharge.

There are other suggested refinements that are crucial for both assessing the current state and measuring the impact of any changes. Chief among these are any process improvements of the type that make it easier for CMs to document the work completed for a case. As discussed, there is strong evidence that higher workload days are associated with more undocumented work. This seems a natural consequence of prioritizing efforts given a limited amount of time to actually perform work and subsequently document this work. In most cases it is easy to infer missing work, though there is greater uncertainty in assessing the magnitude of undocumented work. By extension, an incomplete record of work introduces a lot of noise into both automated retrospective scoring and modeling that depends on these scores. It may be worthwhile to expand or modify the types of templated notes available to case managers for quickly capturing the majority of work completed for a case while allowing transmission of key information to other members of the care team. As a start, some of the non-native note types for which designations were developed in Chapter 4 (e.g., acceptnote, denynote, treport, etc.) could be made more standard. These additional standard note types would have the added benefit of making automated phase/state identification in the record easier, as well as decreasing the error associated with automated retrospective scoring.

It is difficult to gage exactly how serious a problem undocumented work is, but it is enough of a

concern that it played into the decision to use White 8 as the floor to develop the work metric and techniques used over the course of our work. CM leadership identified the primary White 8 CM (responsible for 24 of 26 beds on White 8) as being relatively assiduous in documenting work. In comparison, there seemed to be more undocumented work on White 9, a similar floor to White 8. This discrepancy can likely be traced, in large part, to the fact that White 9 had a much larger number of case managers providing notes for the record examined; often White 9 was staffed by a "float" case manager and this decreased continuity probably had effects on the relative completeness of the record. Even for White 8 there was some concern that the record may have become less complete over time, and particularly in the summer as the primary CM, for example, was on vacation. As indicated in Table 7.1, there did not seem to be any negative correlation between a later date and the amount of work documented on a day; slight, but statistically insignificant positive correlations were identified.

Table 7.1: Correlation between the date and the amount of work scored by day of the week, White 8, 1 October 2014 – 30 June 2015

	Correlation coefficient	*p-value*
Monday	0.114	0.488
Tuesday	0.167	0.329
Wednesday	0.293	0.070
Thursday	0.001	0.997
Friday	0.079	0.632

The work score is a fundamental metric for allowing comparison between floors with different patient populations. The metric facilitates simulation of changes, using historical data, to see how the workload for a CM position is affected. Furthermore, and as discussed in Appendix B, a refined metric may allow for empirically supported and rational MGH-specific caseload benchmarks to be developed; that is, the metric can help determine the number of cases/beds a CM, given a patient population with certain characteristics, should have relative to another CM position responsible for a patient population with different characteristics. However, it is possible to partially decouple setting caseload benchmarks from the work metric, under a set of supportable assumptions, by relying primarily on the number of active days for a CM position over a given interval. "Patient population characteristics" is an admittedly vague concept, but, based the analysis in Chapters 4-6, the most important patient population characteristic, that may extend to floors of types other than that of White 8 and White 9, is how many days a "typical" case is active.

7.2 Staffing patterns: Static (baseline) and dynamic aspects

As set out in Chapter 2, one goal for our work was to:

> Developing candidate staffing schemes that incorporate the flexibility required to effectively address variability in a case manager's daily workload and/or reduce observed variability

The current state of the work precludes the overall level of specificity for staffing recommendations envisioned when beginning our work. However, it is possible to make some specific recommendations and, where specificity is not possible, to show how our work can be extended to develop staffing policies allowing the above goal to be realized.

The general outline of the staffing policies suggested by our work would have both a baseline (static) component and a flexible (dynamic) component. The baseline component is considered more fully in Appendix B, as it is closely tied to supported benchmark caseloads. Even though we showed that the daily census is a poor predictor of the daily workload, this baseline caseload is important to establish. This baseline caseload, or number of beds a case manager is responsible for, if properly set, provides the basis from which a feasible dynamic element can be introduced; any gross deviations from a proper baseline will decrease the effectiveness of dynamic operational aspects.

The proper baseline is also important because there will always be irreducible uncertainty in predicting the daily workload for a case manager because of factors outside of the sphere of influence of the case manager, and not totally attributable to the case characteristics; these factors can greatly impact workload on a daily basis. One example of this would be a case with an anomalously high peak workload. This could occur if conditions align perfectly to facilitate the rapid discharge of a patient, such as delays in the request-response cycle being much less than typical. Proper staffing benchmarks will help to mitigate against these unpredictable effects that are noise from a modeling perspective but that can have very real consequences for a given case manager on a given day. Supported benchmarks also address, in part, CM complaints about an unequal distribution of workload.

The following subsection discusses some preconditions that are necessary for the procedure in Appendix B to have broad applicability. In essence, these preconditions can allow "conversion factors" to be established to equate a typical case on one floor with a typical case on another floor; again, typical only has meaning over long time horizons. This subsection is more qualitative in nature.

The final subsection of this thesis considers the use of pooling, in conjunction with appropriate caseload benchmarks, to provide the basis from which a dynamic element can be introduced. Rather than dynamic staffing a "one-switch" dynamic case assignment scheme is considered. This scheme builds on much of the previous work presented in this thesis and, in particular, the concept of an active census, the phased framework of Chapter 5, and the fact that, for cases active on multiple days, the majority of work occurs in the discharge window. This section requires some assumptions about how CMs experience variability in workload, both upside and downside variability, and why measures that treat upside variability and downside variability symmetrically, such as standard deviation or derived measures like a coefficient of variation (relative standard deviation), may not be appropriate. We examine the use of pooling to make the dynamic case assignment scheme feasible. In our scheme pooling is used to reduce upside variability by attenuating the level of work required of case managers, in a pool, on higher workload days.

7.2.1 Determining baseline staffing: Necessary preconditions

Just as there are two components to staffing, a static and dynamic component, there are components to baseline staffing. One component concerns overall staffing across all floors by day of the week.

The other component concerns the benchmark caseloads across different floors (Appendix B). The latter is the subject of this subsection, but the former deserves at least cursory consideration here and may be a subject worthy of future investigation. To be sure, there is a weekly periodicity observed in the daily workload for case managers. This has been remarked upon and analyzed at various points in this thesis. On the face of it, there may be benefits to be gained by allocating the available CM hours differentially throughout the week. In the simplest incarnation of this differential daily staffing, in general terms, there would be more CM hours allocated for Monday and Friday. However, simulating the effects of such a change is not straightforward. The number of case manager hours to be allocated is finite and allocating hours from Tuesday to Monday, for example, likely results in a delayed work phenomenon, currently evidenced for admit window work on Monday, leading to higher average workloads on Wednesday.

It seems plausible that a differential weekday staffing scheme could result in less cyclical weekly variability in a CM's workload. Again, the exact shape that this scheme would take is difficult to specify, but some general characteristics can be postulated. Case manager coverage would still have to be provided for all floors on all days to help facilitate discharges. A smaller number of case managers, though not as small as on the weekend, would be responsible for a larger number of floors, though, similarly, not as large as on the weekend. The way to investigate the possible benefits of such a differential weekday scheme is straightforward, although more than two floors would have to be used to investigate possible benefits.

For the sake of illustration, consider a pool of four floors with a similar patient population, such as four general medicine floors. During the course of the week these floors would consume 20 CM days. Hypothetically, let us say on Tuesday, Wednesday, and Thursday, these four floors would be covered by, again as an example, three case managers, for a total of 9 CM days consumed. This allows reallocation of 3 CM days to Friday and Monday, so that 11 CM days could be allocated between the four floors on Friday and Monday, an allocation that could be determined, in part, by a prediction of the expected daily workload for the floors.

This general type of differential weekday staffing scheme could result in less observed CM workload variability, but only from a more global perspective. From the individual case manager perspective, of course, variability increases – weekdays with work are interspersed with weekdays without work. This statement is not meant to be facetious, as it is a significant factor to consider for CMs used to a certain schedule and benefits would need to be well supported via simulation to even contemplate undertaking such a change. Scheduling becomes a more complicated endeavor, particularly with unplanned absences, though this is manageable. Potentially more problematic is that some positions would have to either be converted from 8 hours/day to 10 hours/day, and full-time CMs now working only during the week may have to work on the weekend to have a full-time workweek.

There is also the effect of CM discontinuity from the patient's perspective to consider. That is, at least in some cases, it seems likely that the rapport and trust established between a case manager and a patient or patient's family may be important in facilitating the most rapid discharge possible. The lack of continuity may prevent a rapport from developing or, once developed, may make it difficult to leverage. The consequences of a lack of CM continuity for a case are certainly worthy of further study.

Again, there are a number of differential weekday staffing schemes that could be investigated via simulation, but the disruption to long-established CM position schedules will be significant

and a lack of weekday CM continuity may be significant. There are other changes that could be investigated, such as performing more admit window work on the weekend, although it is difficult to tout potential benefits unequivocally. For example, many patients admitted on the weekend (beginning Friday afternoon following normal business hours) discharge later in the weekend or on Monday without ever being screened for high-risk criteria or undergoing an initial assessment. Presumably these patients would fall into the category of LR-Actual patients described in Chapter 6, but if admit window work is directed to be completed on the weekend this could essentially be wasted effort for these patients. Also, though the high-risk screen and initial assessment enjoys a level of standardization, different CMs undoubtedly employ slight modifications in conducting the HRIA. Therefore, an HRIA performed on the weekend would likely result in some rework by weekday CMs, either in the form of a chart review or a brief meeting with the patient on Monday, for example.

Weekly periodicity is certainly one proverbial "elephant in the room" that must be considered when examining one of the components of baseline staffing. However, much more potentially significant is the division of available CM hours between the utilization review function and discharge planning function. Determining the proper allocation of CMs among these two functions is one precondition for a more definitive determination of benchmark caseloads for the discharge planning function. Just as some metric is needed to compare the work done by DCP case managers for the "typical" or average case on different floors to establish benchmark caseloads, so too is a metric to assess whether the current division of labor between UR case managers and DCP case managers is equitable. From interviews it is clear that some CMs, primarily DCP CMs, feel the current division is not equitable, despite the fact that UR CMs have twice the caseload of DCP CMs on average[1]. Whether this is the case needs to be examined in future work to definitively establish the proper caseload for DCP CMs. Establishing the proper division of CM resources between the DCP and UR function is necessary to reveal how many CM resources are available to allocate when establishing DCP CM benchmark caseloads across disparate floors.

Another precondition for determining benchmark caseloads across floors requires a higher level of standardization than observed during the course of this work. There is no need to belabor this point, as it was discussed at some length in Chapter 4, but floors have different role boundaries and processes in which CM processes are embedded. This means that it is possible that the exact same case could require a different amount of CM work on one floor as compared to another. Efforts are underway to examine and standardize care team roles, while eliminating redundancies; such efforts would be beneficial for determining appropriate benchmark caseloads. Relatedly, the role of case management resource specialist (CMRS) personnel needs to be standardized across floors to allow benchmark caseloads to be established. On some floors these personnel are well-integrated into the CM work processes; on others they are not utilized as effectively as possible by the floor CMs. As the work detailed in this thesis was nearing completion, efforts were underway to standardize the role of CMRS personnel across floors. Finally, even the information exchange mechanisms on floors are non-standard. Though we cannot quantify the effect of this type of non-standardization, observation makes it clear that some information exchange mechanisms increase the time required of CMs to "track down" information, held by other members of the care team, required to facilitate discharges.

Admittedly, some of the preconditions for definitive determination of benchmark caseloads are not entirely within the control of case managers and CM leadership. However, the degree to which

[1] Attribution of interview statements withheld.

these preconditions are met increases the confidence that appropriate benchmark caseloads can be determined. As shown in Chapter 6, a level of standardization is also a prerequisite for a widely applicable predictive model. The delay in admit window work for White 9 (delayed HRIA for patients), as compared to White 8 meant that the model incorporating predictors of daily admit window work (or total work) trained using the White 8 data performed poorly when tested on the White 9 data. Of course, it would be possible to train models specifically for different floors, but standardization is key or there will likely be model drift as processes change, over time, at different rates on different floors.

7.2.2 Conditional dynamic assignment of cases within a pooling framework

Pooling is a typical strategy to reduce variability (or risk when risk is measured by variability). As Table 7.2 shows, by commonly used measures, such as the relative standard deviation (coefficient of variation) or quartile coefficient of dispersion, pooling does reduce the variability of the combined daily workload for White 8 and White 9 in Q3 FY15, as compared to the variability for either of the individual floors. This is true for all weekdays on both floors. However, as Figure 7.1 graphically demonstrates, this effect may not be all that significant. In this figure there are two y-axes scaled appropriately for the level of combined workload (left y-axis) and the workload for individual floors (right y-axis). Each data point represents the workload for a weekday in sequence from 1 April 2015 – 30 June 2015. The picture painted by Figure 7.9 is equivocal when it comes to the effect of pooling on total workload variability. In fact, the murky nature of this figure serves as an appropriate metaphor for the unclear effects of pooling on inter-day total workload variability from a practical or pragmatic perspective.

In fact, measures like standard deviation, or measures derived using standard deviation, may not be appropriate for the task at hand. The standard deviation does have some "nice" properties if one makes assumptions about the underlying distributions, but as a mathematical construct it is largely divorced from how CMs are likely to experience inter-day workload variability. This is easy to demonstrate merely by considering the symmetric nature of standard deviation; workloads far above the average are treated the same as workloads far below the average. Analagous objections are sometimes raised to using variance as a measure of risk for portfolio optimization; to an investor upside risk may not experienced the same as downside risk.

A multitude of other measures for the effect of pooling on inter-day workload variability are arguably as valid, if not more so, than standard deviation (or relative standard deviation) or quartile coefficients of dispersion. Table 7.3 lists some of these possible measures, including relative mean absolute deviation, relative semi-deviation for scores above the median, or the coefficient of variation for the active census. Using basic feature scaling, with values constrained to be between 1 and 3 inclusive, corresponding to a low (1), medium (2), or high (3) workload day, also allows other measures capturing the effects of pooling to be developed and considered. The number of consecutive back to back high work days or days above the median could be a valid measure. Of note, pooling does little to decrease the weekly periodicity in workload; if anything the effect, as captured by noting the day of the week corresponding to the thirteen highest workload days for the thirteen weeks in FY15 Q3 is magnified. The count of days falling within a certain number of standard deviations of the average also reveals the equivocal effect of pooling on inter-day workload variability.

The synergy effect noted at the bottom of the table is, however, provocative. This is the effect

Table 7.2: The effect of pooling in reducing inter-day workload variability using customary measures

White 8	All	Mon	Tue	Wed	Thu	Fri
n	64	13	12	13	13	13
Minimum	40.0	59.0	44.0	40.0	48.0	52.0
10th percentile	54.5	59.4	48.2	44.4	48.0	55.6
25th percentile	62.3	63.5	62.3	60.5	58.5	63.5
Median	78.5	83.0	72.5	72.0	79.0	85.0
75th percentile	90.8	94.5	89.8	88.0	88.5	102.0
90th percentile	106.0	105.2	103.2	93.8	108.2	134.0
Maximum	146.0	108.0	108.0	95.0	111.0	146.0
Mean	78.4	81.0	75.0	73.2	77.1	85.5
SD	19.7	16.5	17.7	17.0	19.6	26.4

White 9	All	Mon	Tue	Wed	Thu	Fri
n	64	13	12	13	13	13
Minimum	28.00	38.00	29.00	37.00	28.00	48.00
10th percentile	38.50	43.20	31.40	38.20	32.40	48.40
25th percentile	51.00	53.50	39.50	49.50	51.00	59.00
Median	63.50	59.00	55.00	63.00	64.00	86.00
75th percentile	85.75	77.50	63.75	88.00	85.00	95.00
90th percentile	98.00	109.00	97.00	97.80	94.00	99.60
Maximum	127.00	127.00	109.00	99.00	98.00	100.00
Mean	66.97	67.08	55.50	67.38	64.77	79.23
SD	21.59	21.94	20.62	21.06	21.25	19.52

Pooled	All	Mon	Tue	Wed	Thu	Fri
n	64	13	12	13	13	13
Minimum	91.0	97.0	99.0	91.0	103.0	111.0
10th percentile	111.0	104.2	102.9	93.4	106.2	111.0
25th percentile	127.5	124.5	117.3	125.0	124.5	141.5
Median	145.0	151.0	132.0	147.0	140.0	159.0
75th percentile	160.0	175.5	144.8	162.0	152.5	187.0
90th percentile	184.0	183.8	150.9	176.0	194.2	227.2
Maximum	238.0	187.0	153.0	184.0	209.0	238.0
Mean	145.4	148.1	130.5	140.5	141.8	164.8
SD	29.1	27.7	15.9	26.5	27.2	36.6

	Relative Standard Deviation					
	All	Mon	Tue	Wed	Thu	Fri
n	64	13	12	13	13	13
White 8	0.25	0.20	0.24	0.23	0.25	0.31
White 9	0.32	0.33	0.37	0.31	0.33	0.25
Pooled	0.20	0.19	0.12	0.19	0.19	0.22

	Quartile Coefficient of Dispersion					
	All	Mon	Tue	Wed	Thu	Fri
n	64	13	12	13	13	13
White 8	0.19	0.20	0.18	0.19	0.20	0.23
White 9	0.25	0.18	0.23	0.28	0.25	0.23
Pooled	0.11	0.17	0.10	0.13	0.10	0.14

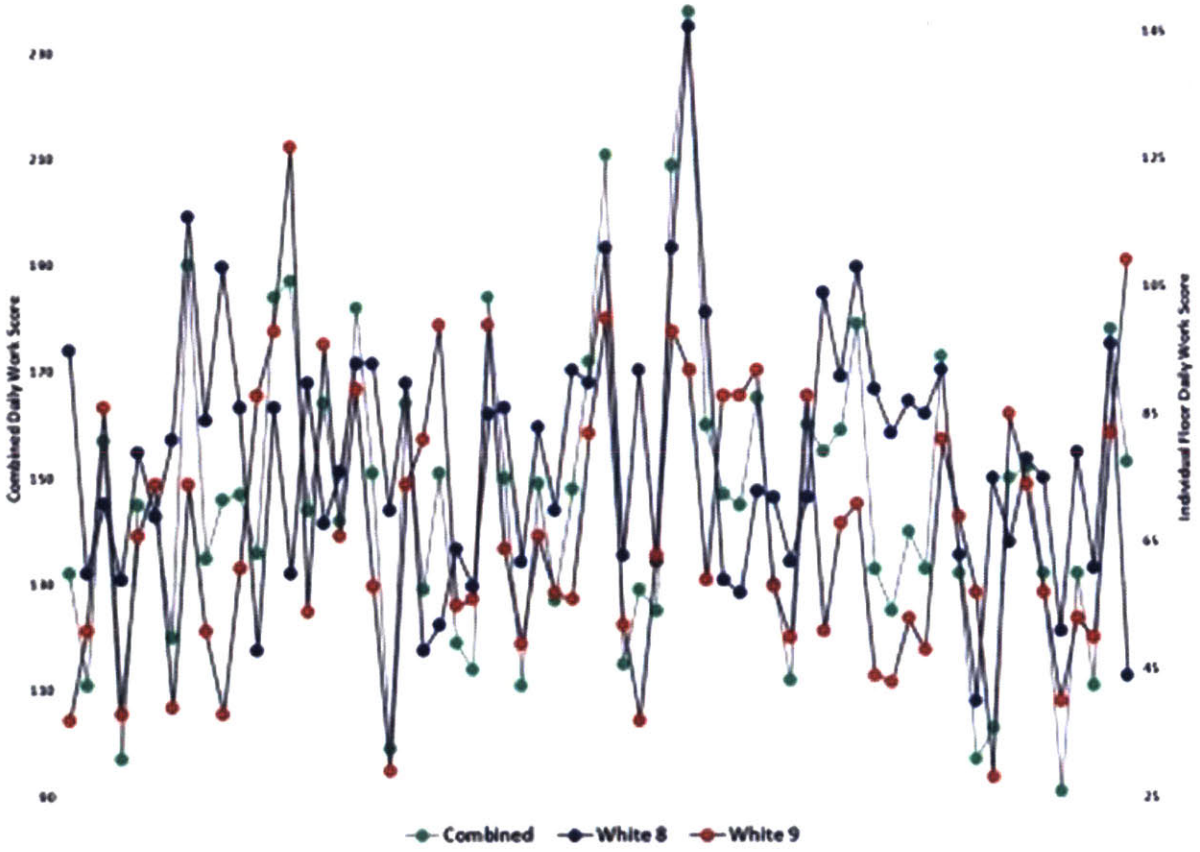


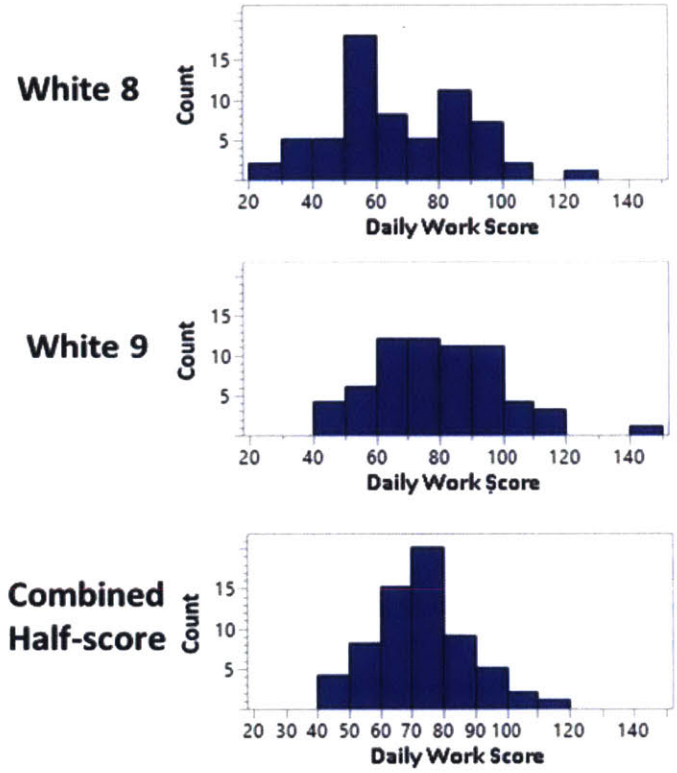
Figure 7-1: The equivocal nature of a pooling effect on inter-day workload variability

Table 7.3: Alternate measures of the potential effects of pooling

	Relative MAD	Relative Semi-Deviation	Relative SD Active Census	Feature-scaled H/M/L Days	Back to Back HW Days	Back to Back Days Above Median	SD Distance Counts < 1 / 1-1.5 / > 1.5	Top 13 Days
White 8	0.20	0.22	0.23	12 / 31 / 21	3	17	48 / 6 / 10	Fri/Mon - 8, Other - 5
White 9	0.27	0.27	0.31	17 / 27 / 20	4	16	43 / 16 / 5	Fri/Mon - 7, Other - 6
Pooled	0.15	0.19	0.21	12 / 34 / 18	4	14	43 / 13 / 8	Fri/Mon - 10, Other - 3

"Synergy" effect of 27 days (16 for White 9, 10 for White 8, 1 for both floors)

whereby a high workload day for White 8 (White 9) and a low workload day for White 9 (White 8) lead to a combined medium (near average) workload day. The effect can also occur when two medium days, or one average and one medium day, leads to a feature-scaled low workload day. This is not a standard metric but it does point to how pooling should be used – not as a mechanism to form the basis for a reduction in total inter-day workload variability, but as a mechanism to allow attenuation of the magnitude of high workload days. Figure 7-2 illustrates this more clearly.



	White 8	White 9	Dynamic Pooling LOP
n	64	64	64
25th percentile	62.3	51.0	63.8
Median	78.5	63.5	72.5
75th percentile	90.8	85.8	80.0
90th percentile	106.0	98.0	92.0
Maximum	146.0	127.0	119.0

Figure 7-2: The use of pooling as a basis for reducing the magnitude of high-workload days (upside variability)

In Figure 7-2 the distribution of daily work scores for White 8 and White 9 is shown, as is the distribution if we consider White 8 and White 9 as a single unit with two CMs available for coverage. In the accompanying table the rightmost column is labeled as the dynamic pooling limit of performance (LOP). This is the amount of scored work that each of the two case managers would have if it could be divided in half. The entries highlighted in green show where the dynamic pooling LOP is less than the quartile cutoff point for the individual floors.

The dynamic pooling LOP is, of course, a theoretical limit. Work is not perfectly divisible between case managers. Yet, it should be possible to approach the dynamic pooling LOP with a "one-switch" dynamic case assignment scheme (1SDCA). The mechanism of the 1SDCA is simple and involves managing both components of the active census, active admit window cases and active discharge window cases. For the first component, on a given day the number of unassessed patients is divided among the two case managers. This is the initial assignment of the case. The "one-switch" occurs, or may occur, when shifting responsibility of the case from one CM to the other as a case enters the discharge window to maintain as close to an equal split as possible of cases in the discharge window.

The performance of the 1SDCA scheme needs to be simulated and this is the next phase of work. The "one-switch" is not cost-free, as there is some work required for a CM to become familiar with a case that is newly assigned as it enters the discharge window on the FDD. Pooling itself is not without a cost because potentially more care team members need to be coordinated with. This means information exchange mechanisms will have to be more fully developed. In theory the notes authored by various members of the care team, and accessible to all members, should be an efficient information exchange conduit, although in practice there is still some improvement that needs to occur in this area.

Though not without costs, the costs are not substantially different than a weekend case manager taking over from a weekday case manager to facilitate a weekend discharge. The mechanism is also elegant in its simplicity and relies primarily on the assumption that cases, when active, require similar amounts of work; high aggregate workload cases are usually just active more days, a feature addressed by the balancing mechanism. Consider, at the median level of daily workload for White 8, the LOP indicates a potential reduction in daily workload on the order of 30 minutes (using the admittedly imprecise imputed time value of 5 minutes per unit work score). At the 75th and 90th percentiles the potential time savings can be estimated as exceeding one hour.

Of course more work needs to be done in terms of assessing the feasibility and simulated performance of the 1SDCA scheme. Yet simple pooling based balancing of active census components should have measurable benefits. There are other benefits in shifting the upper end of the daily workload distribution to the left. This allows more time for CMs to spend with patients; the lack of time to do so was frequently cited as a complaint among case managers. Establishing appropriate benchmark caseloads will make a 1SDCA more effective. Furthermore, the more refined segmentation suggested to improved predictive capabilities can be used to develop a better active census balancing scheme. Finally, an enhanced predictive capability can be used to indicate when any additional available case manager resources should be "flexed" to a pool of 1SDCA floors.

Appendix A

Identifying Probable High Workload Cases: Developing an Improved Screening Tool

Much of our early work during this project stemmed from a reasonable, but incomplete, hypothesis. As explained in Chapter 4, the top decile of patients using our work score accounted for 40% of the total work scored in case manager progress notes. The work completed for these high workload (high aggregate workload) cases is, thus, an important component of the total amount of work completed by case managers and (significantly) disproportionately higher, on a per case basis, than work completed for the remaining 90% of cases.

On an "average" day, of course, this means that 60% of the work completed by case managers is for non-high workload cases. Yet, we hypothesized that the proportion of high workload cases for a case manager position would not be equally distributed across a given time horizon. That is, we hypothesized that on some days a case manager's caseload would contain a greater number of high workload cases. As Figure A-1 shows, this part of our hypothesis turned out to be true.

We further surmised that a caseload containing a greater number of high workload cases should be strongly correlated with an increased daily workload. Furthermore, by tracking the high workload census component of a case manager's workload we believed we could identify a reliable signal for shifting any available case manager resources, such as any reserve case manager capacity or case management resource specialist personnel, to deal with a predicted high workload day. If floors were pooled, our reasoning continued, then it should be possible to balance the number of high workload cases among pooled case managers to decrease the workload variability, in particular extreme fluctuations in upside workload variability, experienced by individual case managers on a daily basis.

As we demonstrate conclusively in Chapters 5 and 6, this chain of reasoning proved to have limited power, both for predicting daily workload and in guiding the development of staffing policies and mechanisms to effectively manage workload variability. Even if one can predict with 100% accuracy on the day of admission that a case will be a high workload case, this prediction says nothing about the timing of the work for high workload cases. That is, the prediction of a high workload case

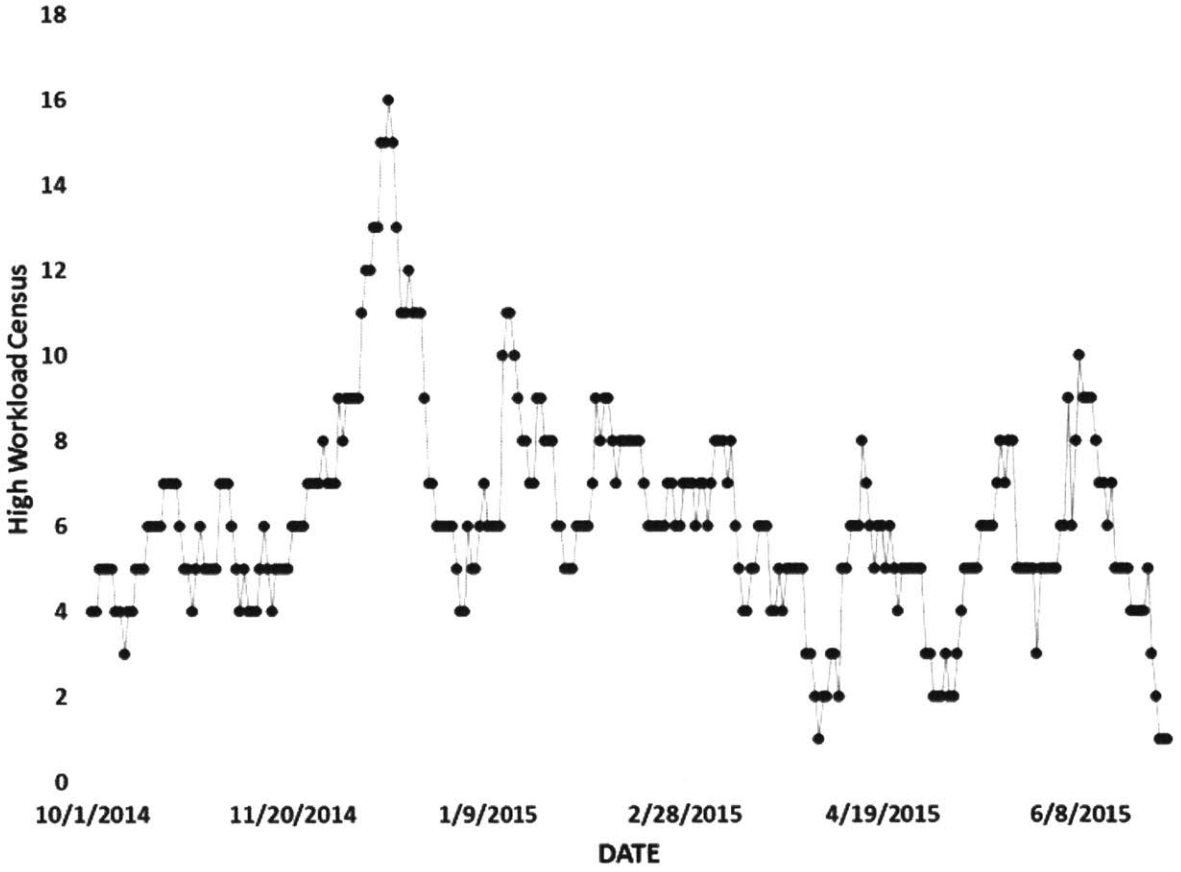


Figure A-1: High workload census on White 8, 1 October 2014 – 30 June 2015 (04:00 census)

refers to a high aggregate workload over the course of a patient's LOS.

More effective than differentiating cases comprising a caseload on the basis of predicted workload over the course of a LOS was differentiating cases as active or inactive on a given day, and then further differentiating active cases using the phased framework in Chapters 5 and 6. This scheme of differentiation also allows development of staffing schemes introducing a dynamic element, such as the 1SDCA scheme in Chapter 7. Rather than balancing the number of high workload cases between case managers, which may or may not be active on any given day, we propose balancing the number of cases by phase in the discharge planning plane and show how this balancing is more effective for managing workload variability.

In some respects our early work can be understood as an attempt to develop a better high-risk screen and initial assessment (HRIA). As shown in Chapter 6, from a negative predictive value standpoint, the current HRIA is noteworthy – if a patient either fails to meet high-risk criteria or, upon initial assessment, the case manager determines no intervention is required, the case has less than a 3% chance of being high workload. However, the false positive rate is very high – many patients that either meet high-risk criteria or who cannot be definitively ruled out as requiring case manager intervention ultimately require little work from a case manager. Our initial work sought to develop a classification tool that maintained the sensitivity of the current HRIA (identifying high workload cases) while also increasing the specificity (decreasing the number of false positives). The results of classification would then be used as an input to our predictive model for daily workload.

As discussed above, and in great detail in Chapters 5 and 6, a refined HRIA has limited utility for facilitating accurate prediction of the daily workload for a case manager. However, a refined tool of this sort could identify patients where more focused early intervention could be beneficial. Within the CM department at MGH two case manager clinical specialists are tasked with assisting on especially difficult cases. From interviews with the specialists it is clear that early identification and subsequent intervention can have a significant impact on the total amount of work that must be completed for these cases[9]. The current state of our work does not allow quantification of the benefits of early intervention in difficult cases, but this is likely a valuable area of future research. Based on our exchanges with case manager specialists and the case manager leadership team we speculate that an improved HRIA tool would need to focus not on the top decile of cases by workload, but on the top 2%-5% of cases by workload for floors similar to White 8. As an anecdotal aside, preliminary work in this area indicates these extreme high workload cases can be identified with a much lower false positive rate, as compared to top decile cases by aggregate workload, and a sensitivity in excess of 97%.

Even if an improved HRIA has limited utility for predicting daily workload, the explanatory power from this line of research could result in significant benefits for the case management department at MGH. For example, if end-of-life (EOL) issues prove to have significant predictive and explanatory power when looking at the total work associated with a case, then results of this sort could indicate the need for (i) specific types of training for case managers at MGH, (ii) a modification of the MGH dyad model of case management that more clearly codifies and delineates the role of social workers, or (iii) even be used as support for certain types of reorganization[123]. Though this is just an example, the point should be clear – identifying which combination of factors lead to high workloads for a cases can inform decisions about requisite training for case managers and role boundaries.

Our early work was also beneficial in developing the techniques used to effectively cope with

imbalanced learning issues arising from many of our sub-problem formulations. Our use of techniques like SMOTE(see Chapters 3 and 6) were refined in the context of developing an improved HRIA. Figure A-2 and Table A.1 provide a flavor of the suite of machine learning and sampling techniques employed during this phase of our work. Though a full discussion of Figure A-2 and Table A.1 is beyond the scope of this thesis, it bears noting that a multi-step feature selection process was key to this phase of our work in order to avoid overfitting training set data. It should also be emphasized that increasing the size of the data set does not lessen the necessity of judicious selection of features for inclusion in an explanatory or predictive model[1]. Our imbalanced learning problem, coupled with the desire to achieve high sensitivity and specificity (or balanced accuracy) means that, relatively independent of the sample size, research efforts will always have to contend with issues more typical of high p (dimensions or features), low n (sample size) problems.

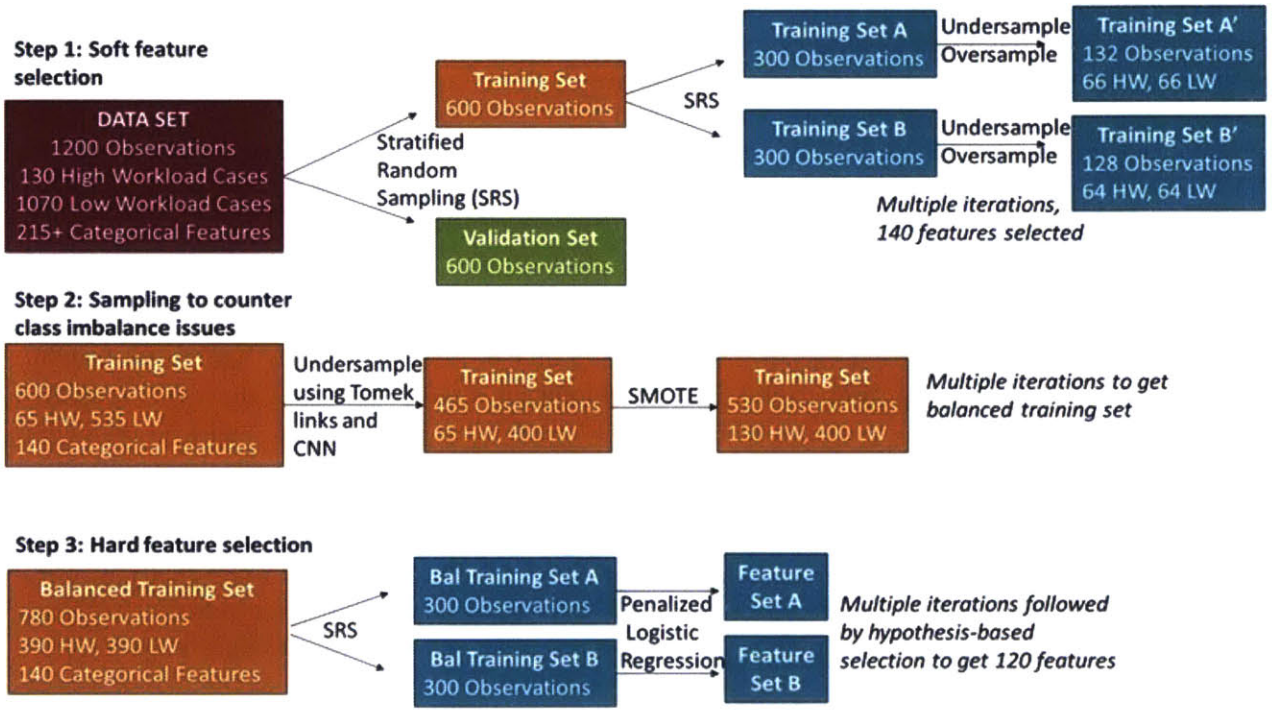


Figure A-2: Overview of sampling and machine learning techniques used for feature selection when predicting high workload cases

We are able to offer some words of caution for future researchers developing retrospective explanatory models for case features and feature interactions driving total workload for a case. Our work, just as other work reviewed in Chapter 3, indicates that psycho-social factors, more so than clinical factors, are key determinants of the work associated with a case. In the data sets available, the presence or absence of these factors was determined by an examination of case managers, particularly HRIA text. However, there is evidence that the prevalence of factors we identified as key indicators for high case manager workload were under-reported, in case manager notes, for low aggregate workload cases. This under-reporting has the effect of making certain factors seem more indicative of a high workload case than if the same type information was routinely recorded for every case. As indicated

[1]This statement holds even if algorithms with inherent feature selection mechanisms are used, such as penalized logistic regression or classification trees. The crux of the issue is that high workload cases can be as different from each other, in feature space, as they are from low workload cases; this greatly increases the chances of overfitting a training data set unless intelligent sampling techniques are used to train a model with a more general bias.

Table A.1: Overview of steps used for training and testing classifier used in predicting high workload cases

	STEP	TECHNIQUES	RESULT
1	"Soft" feature selection	Undersampling, Oversampling, Fisher's Exact Test for Count Data, Odds Ratio	Select 140 out of 215 features
2	Create balanced training set	Tomek links, Condensed nearest neighbors, SMOTE	Create balanced training set with 260 high workload observations (195 synthetic), 260 low workload observations
3	"Hard" feature selection	Penalized (LASSO) logistic regression, Fisher's Exact Test for Count Data, Odds Ratio	Select 129 out of 140 features
4	Hypothesis-based feature se	Penalized (LASSO) logistic regression	Select 120 out of 129 features
5	Train and test model	C5.0 decision tree, adaptive boosting, misclassification costs	Four models applicable to different point in time during a patient's stay (based on features typically revealed by this point)

by the negative predictive value of the current HRIA, when case managers note that a case is unlikely to require a high-level of case manager intervention they are usually right. However, once case managers make this determination then less information is likely to be recorded for a case.

As an example, cognitive deficits or behavioral problems seem to be indicative of high workload cases. However, these factors are only indicative of high workload cases in combination with, or interaction with, other factors, such as the need to be placed in a sub-acute facility upon discharge from MGH. When case managers review a case, if cognitive deficits or behavioral problems exist but it is clear to the case manager that the patient will be safe to discharge home then these case factors may not be indicated in the case notes. Similarly, as a case progresses and more notes are written, factors often present from the outset of a case are eventually documented as an explanation for why a patient is difficult to discharge. In a reading of the notes these complicating factors may appear as emergent when, in fact, they were present at the outset. In some cases other data sources, such as those briefly discussed at the conclusion of this appendix, can be used to complete the feature set for a case. However, it seems highly probable that features are under-reported for low workload cases.

One solution to this problem, from a data completeness standpoint, would be requiring every field of the HRIA shown in Appendix D to be documented. From a workload/operational standpoint this is not acceptable as it fails to leverage the observed experience and judgement of at least some case managers in identifying cases not requiring case manager intervention. This would greatly increase the documentation workload of case managers and the value, having more complete data for building explanatory models, while of interest for researchers, likely does not justify the cost. This is one reason why some documentation requirements have been relaxed (e.g., there is no requirement for an HRIA of patients admitted and discharging over the weekend). From a predictive standpoint improvements to our model outlined in Chapters 6 and 7 should allow leveraging of the negative predictive value of the current HRIA. Furthermore, it may be possible to construct a more complete data set of case features from other sources, such as clinical notes and consult notes/orders not written by case managers.

A final word of caution concerns codifying either binary features or the definitions for features with three or more levels. For example, what is the definition of a cognitive deficit? Does this mean only some type of organic dementia or does it include short-term memory loss or traumatic brain injuries? Similar questions could be asked about behavioral problems or a decreased cooperation feature. How many levels are needed for behavioral problems? Does a patient requiring mechanical or chemical constraints, exhibiting documented non-compliance with a treatment program, or having a propensity for leaving the hospital against medical advice get coded as a behavioral problem or are these better considered as distinct categories? If distinct categories, given that interactions among case features are more important than any main effects, a very large data set is required. These types of categories, even if coding for modeling suggests an objective factor, are often rooted in subjective evaluations by members of the care team that may vary between practitioners. This is to say nothing of the fact that many case features depend on self-reporting by the patient or family, another potential source of under-reporting.

The phased framework of Chapter 5 eliminates many of the problems that an incomplete case feature set can introduce; our conventions introduce less ambiguity into the data set used for modeling. This approach facilitates predictive modeling of daily case manager workload, but at the cost of not being able to identify, at a granular level, specific case features impacting workload for a given case.

We were primarily concerned with an improved HRIA for the purpose of identifying a high aggregate workload case as early in a patient's LOS as possible. As discussed, this type of early identification at the individual case level is not required for prediction of a case manager's daily workload, nor is it necessary to inform development of staffing schemes that decrease the more onerous effects of daily workload variability. In our early work we identified four general points (or periods in the case of 3 and 4 below) during a patient's LOS when the prediction of a high aggregate workload case could be made:

1. Before any HRIA by the case manager
2. Immediately following the HRIA
3. Following the HRIA but no later than the first active discharge window day
4. After the first active discharge window day up to discharge day

These points were distinguished by the information usually available at each point. For example, at point 1, basic data would be available, such as the age of the patient and where/what type of facility the patient was admitted from. The **predictive** identification of a high aggregate workload case becomes problematic, especially after point 2, because information discovery at a later point means some work for a case has already been completed.

The sample size was 1200 patients discharging from White 8 between 1 October 2014 and 30 June 2015. Figure A-3 shows some of the 215+ case features, typically available at each point, which were identified as potentially important indicators of a high workload case[2]. The scale was developed, primarily for illustration purposes, by constructing an odds ratio for a case being high workload if a feature was present as compared to if a feature was absent. The top decile of cases by our work score were designated high workload cases.

[2]The SCSX scores are described in the main text. Other features shown that may not be recognizable as encoded include FTT (adult failure to thrive), IVABXDuring (patient require IV antibiotics during his LOS), and NotIndADL (not independent with the activities of daily living such as feeding, bathing, and grooming).

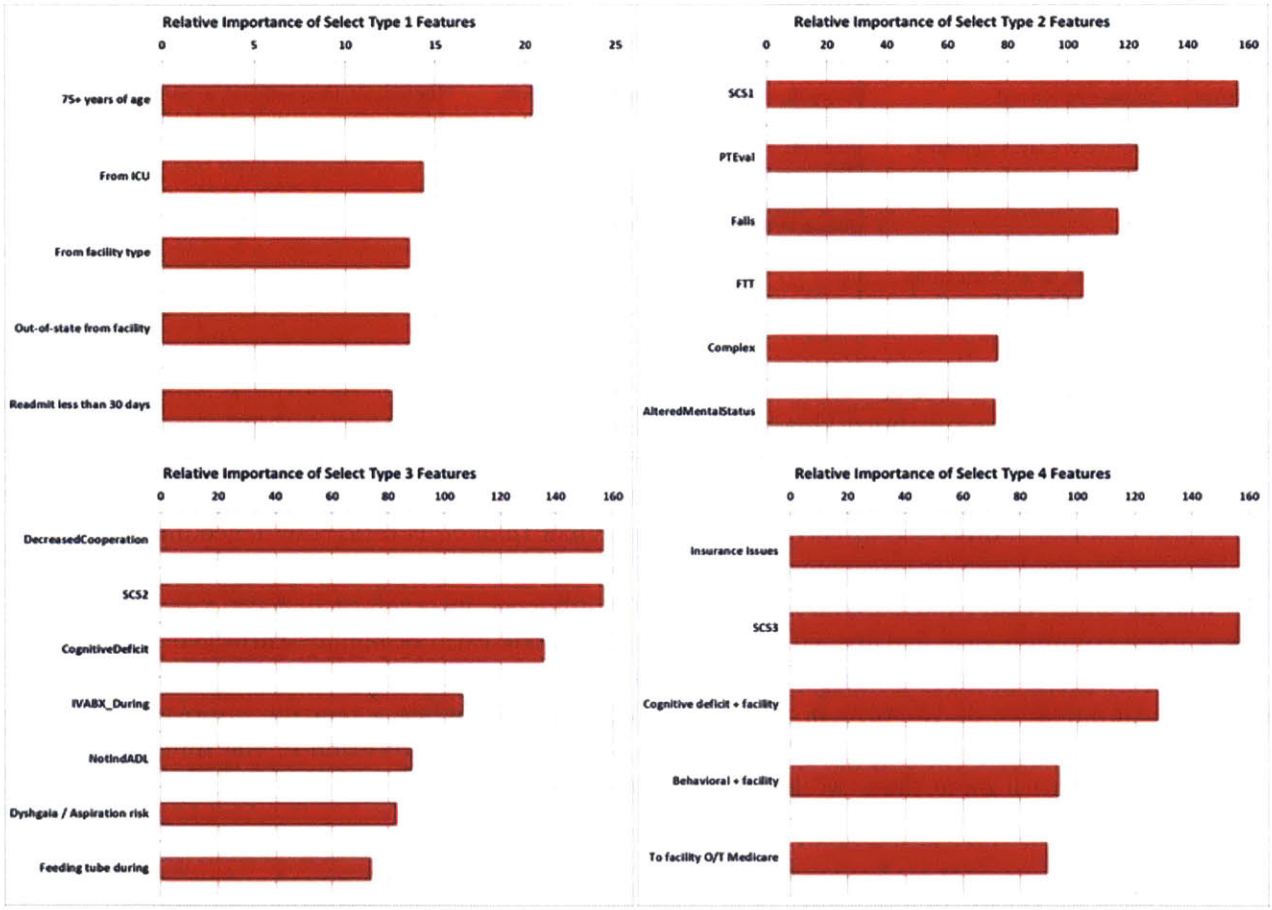


Figure A-3: Relative importance of select case features "typically" revealed at different points of a patient's LOS

Table A.2: Composition of social complexity scores

SCS3	SCS2	SCS1
iCMP	iCMP	iCMP
Social work order (initiated by any)	Social work order (initiated by CM)	Social work order (initiated by CM)
Lack social supports	Insurance issues (initial assessment)	Insurance issues (initial assessment)
Homeless	Homeless	Homeless
Not independent iADL and ADL	Not independent iADL and ADL (initial assessmer	Presents with altered mental status
Cognitive deficit	Cognitive deficit	Substance abuse
Guardianship/legal issues	Guardianship/legal issues	FTT
Substance abuse	Substance abuse	Psych history (initial assessment)
Insurance issues (initial assessmentor emergent)	Decreased cooperation	Diabled (<65 years old)
Decreased cooperation	Behavioral issues	More than two admissions last six months
Behavioral issues	FTT	Transfer from facility
End of life	Psych history (initial assessment)	Readmit < 30 days
FTT (Failure to thrive)	Diabled (<65 years old)	75+ years of age
Psych history (initial assessment)	More than two admissions last six months	Uninsured, international, or worker's comp
Diabled (<65 years old)	Transfer from facility	Active services at time of admission
More than two admissions last six months	Readmit < 30 days	
Transfer from facility	75+ years of age	
Transfer to facility	Uninsured, international, or worker's comp	
Readmit < 30 days	Active services at time of admission	
75+ years of age		
Uninsured, international, or worker's comp		
Active services at time of admission		

For binary features, Fisher's exact test was used to test the null hypothesis that the odds ratio[41][31][88] of being a high workload case, with a given case feature and without the case feature, was equal to one. A measure of relative importance was created as -10*log10(p-value) from this test; this is the scale shown in Figure C.4[88]. This relative measure is more representative of feature importance than the odds ratio directly, and more succinct than odds ratio confidence intervals. The p-value is influenced by the prevalence of the case feature in our sample – a very rare case feature with a large odds ratio (and correspondingly wide confidence intervals) will have a larger p-value than a more common case feature with a more modest odds ratio of higher statistical significance (smaller p-value). In the case of the latter the smaller p-value results in a larger relative importance. A relative importance of 13 corresponds to a p-value of 0.05, while a relative importance of 157 corresponds to a p-value less than 2.2E-16; small p-values lead to rejection of the null hypothesis. For case features with more than two levels, a chi-squared approximation was used in lieu of Fisher's exact test.

As stated, a complete discussion of each of the factors in Figure A-3 is beyond the scope of the thesis. The takeaway from the figure is that some case features identifiable in the available data sources are associated with statistically significant increased odds of association with a high workload case. The "+" factors shown (e.g., cognitive deficit + facility) indicate the case had some feature and admission to a post-MGH facility was required.

One type of factor deserving further explanation is the SCS factor, a social complexity score. We created this score as the unweighted sum of the number of binary features shown in Table A.2 present for a case. The different SCS scores (SCS1, SCS2, and SCS3) were based on features typically available at the different points during a patient's LOS. These scores do not consider interactions among factors, but even the unweighted sums provide a statistically significant indication of a high workload case. Each score had four levels based on the quartile demarcations for the sums. Many of the factors in the social complexity scores reflect the high risk criteria currently screened for.

Table A.3: The absolute prevalence of some features among high and low workload cases

		Low Work	High Work
Insurance Issues Initial Assessment	N	**1012**	**137**
	Y	**36**	**15**

		Low Work	High Work
Insurance Issues Emergent	N	**994**	**84**
	Y	**54**	**68**

		Low Work	High Work
Cog+Beh+Psych	N	**1043**	**145**
	Y	**5**	**7**

		Low Work	High Work
Cog+Beh+Psych+Fac	N	**1043**	**145**
	Y	**2**	**7**

The class imbalance problem means that even a case feature that increases the odds of a high workload case still yields a high level of false positives. For every feature investigated, with the exception of legal/guardianship issues, there were more total low workload cases than high workload cases with that feature. One way to address these problems is by placing more emphasis on the interactions between combinations of case features. Table A.3 demonstrates some important points about our efforts to develop a more refined HRIA.

In the first entry in the table, the odds ratio for being a high workload case when insurance issues are identified in the initial assessment as compared to when these issues are not present is a statistically significant (α=0.05) 3.1. Yet, there are 2.4 times as many low workload cases as high workload cases with insurance issues identified during the initial assessment. Identification of high workload cases improves as the LOS of a patient progresses, as indicated by the second entry in the figure. When we consider emergent insurance issues identified during active discharge planning the odds ratio increases to 14.9 and a greater number of high workload cases exhibit the emergent insurance issue feature than low workload cases.

As shown in entries three and four, combinations of factors may allow even better discrimination between likely high workload and low workload cases. If a patient has a cognitive deficit, a behavioral problem, and a history of psychiatric problems the odds ratio is 10.1; if this patient also has to go to a facility the odds ratio of being a high workload case increases to 25.2. These odds ratios are statistically significant (α=0.05). Of course, entries two, three, and four in the table mean some work has already been completed for a case. Also, as the number of features in a combination is increased there are fewer cases exhibiting the interactions; this makes it difficult to train a classifier without a large sample size or a training data set with carefully simulated data.

Our method to deal with the small sample size revolved around the feature selection algorithm presented in Figure A-2 and Table A.1, and extensive use of the SMOTE technique to train a boosted tree classifier. The results of classification, on an unadulterated test set, at each of three

Table A.4: Performance of boosted classification trees on test sets at different point in a patient's LOS

POINT 4

		ACTUAL CLASS	
		Low Workload	High Workload
PREDICTED CLASS	Low Workload	409	12
	High Workload	10	49

POINT 2+

		ACTUAL CLASS	
		Low Workload	High Workload
PREDICTED CLASS	Low Workload	401	16
	High Workload	18	45

POINT 1

		ACTUAL CLASS	
		Low Workload	High Workload
PREDICTED CLASS	Low Workload	385	31
	High Workload	34	30

"points" during an LOS (distinguishable by the case features typically revealed at that point) are shown below in Table A.4. The point 2+ designation indicates the use of features extracted from data sources other than case manager notes (see below); these features are sometimes available before the initial assessment and sometimes during period 3 so "typically revealed" is a difficult description to support for these features without further analysis.

Rather than reporting performance metrics for these classifiers (as in Chapter 6) we consider the performance more generally, in comparison to the current HRIA. Point 4 and Point 2+ classification correctly identifies 80% and 73% of high workload cases, respectively, with false positive rates below 5%. The existing HRIA, as discussed in Chapter 6, identifies 87%-93% of cases that will require a high workload, but 79%-86% of the total cases flagged with the current HRIA are "low" workload cases (cases not in the top decile of cases by our work score), a substantial false positive rate.

The performance of the Point 1 classifier pales by comparison, identifying just under 50% of the high workload cases, with an 8.1% false positive rate. However, the performance of this classifier is actually quite remarkable given the limited number of features used for classification – only basic features available before an HRIA was completed were included. If the feature set were expanded using an automated search of available health records, including searching for persistent features (such as a permanent cognitive deficit or a history of falls), this type of classifier could prove useful for automated pre-screening that eliminates some of the work for case managers. The problem with the Point 4 and Point 2+ classification is that it occurs later in the LOS. However, these classifiers could likely be improved with future work to form the basis for explanatory models as discussed

previously in this appendix.

In addition to the data sources identified in Chapter 2 that were key for the work described in the chapters-proper of this thesis, other data sources were utilized in our attempts to identify/predict high aggregate workload cases. These sources of data include those derived from the physician order entry system. Among the data available from this system are the date and types of consults ordered, reasons for the consult, discharge orders, restraint orders and reason, diagnostic studies ordered for a patient, and tubes required by the patient. Obviously this is important clinical information but the notes also contain information that supplements the psychosocial factors/features for a patient revealed by case manager notes. The inclusion of certain consults in the record proved to be a useful case feature in our early work on predicting high workload cases. Certain consults, such as physical therapy, may also be useful as a leading indicator for a patient's entry into the discharge window (FDD).

The suitability of certain consults as proxies or high-level indicators for a patient's psychosocial profile was also examined during the course of the work described in this appendix. These consults included social work, addictions, and psychiatric consults, as well as information contained in restraint orders. MCCM data was also used in the work aimed at predicting high aggregate workload patients. This data indicates which patients are iCMP patients and which patients are homeless, both of which may be important psychosocial factors for case managers to consider.

In spite of the caveats with the data sets we used in an attempt to develop a "better" HRIA, this line of inquiry has value. Some of the problems with incomplete data can be overcome by leveraging other data sources, primarily the actual text of notes written by other members of the care team. Also, in many of the cases we examined, patients had case manager case notes from previous admissions; these notes could be mined for retrospective analysis purposes and for assisting a case manager in completion of initial screening/assessment work. That is, it may be possible to automatically screen patients for the presence of both persistent features and features that necessitated admission. In the case of the latter admission documentation for a patient could conceivably be automatically analyzed (text-analytics) to facilitate assessment.

Appendix B

Refining Relative Caseload Benchmarks with a Case Manager-Specific Case Mix Index(CMI)

The paramount importance of the timing of work associated with a case, in terms of predicting the daily workload for a CM position, is introduced in Chapter 5 and discussed extensively in Chapter 6. This importance is encapsulated in the concept of an active census[1].

The active census concept was extended further by designating a case as an active admit window case or an active discharge window case based on the nature of work completed for a case on a given day. The HRIA is clearly admit window work, while meeting with a patient to, for example, get post-MGH facility choices and place a subsequent referral, is clearly discharge window work. The demarcation between admit window work and discharge window work is not always as stark; some cases (described as "combo" cases) have admit window work and discharge window work documented in the same note. With these cases it is possible to apportion the documented work into admit window and discharge window work when manually reviewing cases; i.e., reading through case notes. Automatic apportionment of work components for "combo" cases is not always straightforward.

The concept of an active census is a powerful concept, both for predicting daily workload and as a basis for cross-floor comparisons if certain assumptions hold. The power of this concept is augmented when overlayed with the phased framework developed in Chapter 5, distinguishing between active admit window cases and active discharge window cases. As described in Chapters 4 and 5, the note types, both native header notes and non-native header notes, mark the transition to an active discharge window case.

For floors in which similar types of work are completed in the admit and discharge windows it is possible to reasonably compare the work done between floors over a given time horizon by a simple

[1] As explained in the preceding chapters, the designation of "active" is a minor misnomer because there is often some low level of work associated with a case on a daily basis, such as discussing a case in rounds. The work score for a case does not account for this type of work.

count of total active admit window days and total active discharge days. The underlying assumption is that the number of active days drives the workload for a CM, daily or over any interval chosen. This assumption can be extended by assuming cases have one active admit day and a variable number of countable active discharge days in the record.

The utility of the active census concept can be demonstrated easily when considering floors with ostensibly similar patient populations like White 8 and White 9. Of course, "ostensibly similar" is another assumption that may not hold at all times. For example, when considering the total amount of work scored for White 8 and White 9 in Chapter 5, there was less work (scaled) than expected for White 9, attributable in large measure to the disproportionately high number of "zero" workload cases and 'home with no services' (HNOS) cases on White 9, compared to White 8, during Q3 of FY 2015.

Yet, these types of differences do not matter greatly when using the active census count for comparisons between similar floors. For White 9, the HNOS cases and "zero" workload cases would not contribute to the count of active discharge days over the interval examined. The similarities (or differences) that matter more include the activity ratios (Chapter 5), the aggregate amount of time cases spend in each phase (Chapter 5), the percentage of total scored work completed in each phase (Chapter 6), and even the contributions to total scored work by high, medium, and low aggregate workload cases (Chapter 4). On these measures White 8 and White 9 were shown to be remarkably similar, as should be expected. Table B.1 documents further similarities, by considering the activity ratios; the number of active, active discharge, and inactive days as a percent of the total number of such days; and the average number of active, active discharge, and inactive days for high, medium, and low aggregate workload cases on the two floors.

The agreement between the floors is striking. The only marginally statistically significant difference (α= 0.10) is with the average number of inactive days for high aggregate workload cases (top decile by work score as the convention was presented in Chapter 4).

Table B.2 reveals further similarities between White 8 and White 9. This table shows the variability, as measured by the standard deviation, in the number of active and inactive days for different aggregate workload case groups. Though not important for the discussion at hand, implicit in an examination of these tables is the fact that, given the greater number of inactive days compared to active days, the timing of work, as captured by the concept of an active census, must be exploited to accurately predict daily workload.

Taken in totality, the preceding discussion in this section, as well as the identified similarities in earlier chapters, the assumption that White 8 and White 9 are the same in dimensions that matter from a CM work perspective is well-established. The similarities between White 8 and White 9 help facilitate inter-floor comparisons of the total amount of work completed over a specified time horizon. In fact, the number of active discharge window days can form the basis for a CM-specific measure analogous to the case mix index (CMI).

While a discussion of the CMI is beyond the scope of this work, the basic concept and uses of the CMI is simple to grasp[26][47][17][23][36]. The CMI attempts to capture the average complexity of cases on a floor of the hospital, or in an entire hospital. This complexity is most closely associated with the clinical complexity of cases, as measured by the amount of hospital resources required by a patient during their hospital stay. Each patient has a weight assigned, retrospectively, based on the diagnostic resource group (DRG) that is coded for the patient. This DRG weight considers

Table B.1: Comparing the activity ratio and active/inactive days for high, medium, and low aggregate workload cases on White 8 and White 9

		n	Active Days	Active Discharge Days	Inactive Days	Activity Ratio
Aggregate Counts						
White 8	Low Workload	737	737	178	1969	27%
	Medium Workolad	369	1020	703	1448	41%
	High Workload	123	737	606	976	43%
						36%
White 9	Low Workload	235	222	36	679	25%
	Medium Workolad	117	346	218	470	42%
	High Workload	40	249	194	292	46%
						36%
Percent of Column Total						
White 8	Low Workload		30%	12%	45%	
	Medium Workolad		41%	47%	33%	
	High Workload		30%	41%	22%	
White 9	Low Workload		27%	8%	47%	
	Medium Workolad		42%	49%	33%	
	High Workload		30%	43%	20%	
Average Days						
White 8	Low Workload		1.00	0.24	2.67	
	Medium Workolad		2.76	1.91	3.92	
	High Workload		5.99	4.93	7.93	
White 9	Low Workload		0.94	0.15	2.89	
	Medium Workolad		2.96	1.86	4.02	
	High Workload		6.23	4.85	7.30	

Table B.2: Standard deviation of active and inactive days for high, medium, and low aggregate workload cases on White 8 and White 9

	White 8		***White 9***	
Workload	***Active Days***	***Inactive Days***	***Active Days***	***Inactive Days***
High	2.92	9.07	3.13	9.07
Medium	4.08	0.96	6.02	0.96
Low	3.55	0.63	3.08	0.65

Table B.3: Correlation of various DRG weights and the "Plus one" count with case work score

	DRG Weights					
	APR20	AP21BC	AP21NY	APR26	MS	Plus one
All cases (1176)	0.29	0.24	0.28	0.29	0.27	0.92
All non-zero cases (1113)	0.27	0.24	0.27	0.28	0.26	0.92

diagnosis codes, procedure codes, discharge disposition, age, gender, and other factors. Each of these components may similarly be composed of a number of factors; for example, diagnosis codes consider complications and comorbidities (CC) or major complications and comorbidities (MCC). There are a number of different DRG schemes, and coding diagnoses and procedures correctly, and in the correct order, takes a certain amount of skill to ensure proper reimbursement, based in part (sometimes large part), on the DRG and associated weight[16].

Again, this is just a brief overview of a potentially complicated procedure done by DRG coders that glosses over details not relevant to this discussion. The end result, however, is a DRG weight. The baseline weight for a patient is 1.0; any weight less than 1.0 is meant to represent less than average hospital resource utilization while a weight greater than 1.0 represents greater than average resource utilization. For example, consider the Medicare Severity (MS) DRGs. A lung transplant has an MS-DRG weight of 9.3550, pneumonia (no CC/MCC) a relative weight of 0.7906, and chronic obstructive pulmonary disease with an MCC a relative weight of 1.1924[16]. The CMI is an average of the DRG weights for all patients over a given interval.

Now, there is an assumption that the CM work required by a case may not be adequately reflected in the CMI. This is a reasonable assumption considering the number of psycho-social factors, not necessarily captured by a clinical-complexity based weight, that impact on the CM workload for a case. In fact, as Table B.3 shows, the correlation coefficient between the various DRGs tracked for MGH patients and the work we scored for cases is low. The rightmost entries for these columns, designated "Plus one", show the correlation with a count of a case's active discharge window days plus one day (assumed one active admit window day). The cases considered were those in which patients had White 8 or White 9 as the only inpatient department where CM work was documented or could be inferred. The correlation is shown including and excluding "zero" workload cases.

As an aside, as shown in Table B.4 the correlation between the various DRG weights for a case are not always strong. Lack of perfect correlation is expected as the various DRG weights are calculated differently to consider, exclude, or emphasize certain elements for a patient. For example, AP-DRGs include a more granular DRG coding scheme for non-Medicare patients, especially newborns and children, while APR-DRGs incorporate more refined measures for severity of illness and mortality risk in addition to resource utilization[?][36]. Still, the correlations should give one pause when considering when and how to make use of the DRGs available in the MGH data sets.

The series of scatterplots in Figure B-5 make the consequences of the correlations detailed in Table B-3 clearer from the CM perspective. The APR26 DRG weight was chosen as the abscissa in panel 1 because of the larger, but still low, positive correlation with the scored work. The agreement between the "Plus one" count and the scored work would, of course, be made even stronger by explicit incorporation of the higher work associated with an active discharge window case as compared to an active admit window case. Clearly, the "Plus one" measure could form a valid basis for a CM-specific CMI for similar floors. This is not particularly useful in itself but there is a way to

Table B.4: Correlations between DRG weights for the cases examined

	APR20	AP21BC	AP21NY	APR26	MS
APR20	1.00	-	-	-	-
AP21BC	0.63	1.00	-	-	-
AP21NY	0.62	0.92	1.00	-	-
APR26	0.90	0.65	0.63	1.00	-
MS	0.68	0.76	0.72	0.68	1.00

make comparisons between floors with seemingly disparate patient populations. One method would use the work score to get an average value for an active discharge day for different floors and a count of active discharge days in the record. Yet, this may not be necessary to make comparisons with the goal of establishing benchmark caseloads across different floors.

Consider an oncology floor, Lunder 9, in comparison to White 8. Medical complexity may be a larger determinant in the amount of work required of a Lunder 9 CM for a given case. However, if the consequence of this medical complexity is more active discharge window days rather than more work per active discharge window day, then the count-based method is still valid. For the purpose of establishing benchmark caseloads across floors it is plausible that if Lunder 9 cases have more active discharge window days, often because of shifting discharge disposition targets, this fact is more immediately significant for staffing decisions than why Lunder 9 cases have more active discharge window days[5][2]. This is plausible, in part, because of the limitation frequently imposed on the amount of work that can be completed for a case on any given day because of the delays introduced by the request-response cycle. Furthermore, target discharge dispositions may be characterized by change on a day-to-day basis, rather than an intra-day basis. However, accurately predicting daily workload would still have to consider why a case is (still) in the discharge window. Thus establishing caseload benchmarks, while related to the work score used to predict daily workloads, may be a different problem, and simpler, than predicting daily workload.

Two facts bear noting. First, the note count in the record cannot be used in all cases to get an accurate count of the active discharge window days for a patient. As described in Chapter 4, this is primarily because of missing notes that have to be inferred and notes that document multiple days of work. Still, the concept of an active census, while not sufficient for predicting workload on a daily basis, may be an important concept for establishing benchmark caseloads between floors. Different types of floors need to be examined to confirm or deny this in future work. Second, a measure such as the "Plus one" count, or even the work metric itself, depending on how modeling occurs, could suffer from issues related to endogeneity or, more problematic, omitted variable bias. Chief among the omitted variables could be the skill/experience of the case manager in handling cases with specific attributes. Some case managers may be more proficient at expeditious facilitation of discharges of a certain type than either other types or in comparison with other case managers for a given type. Too, the work and number of active discharge days for a patient could depend on factors not completely attributable to the patient or factors outside of a case manager's sphere of influence.

[2]During at least three meetings with CM leadership one of the nursing directors made the statement that more than 10% of the patients on an oncology floor would be classified as high workload. The reasoning behind this statement was that discharge dispositions change often for oncology patients. If this is true then the predictive model developed based on examination of cases on general medicine floors should still be extensible to oncology floors. What is more, a measure like "Plus one" would allow meaningful comparisons between, for example, White 8 (general medicine) and Lunder 9 (oncology) over a specified time horizon.

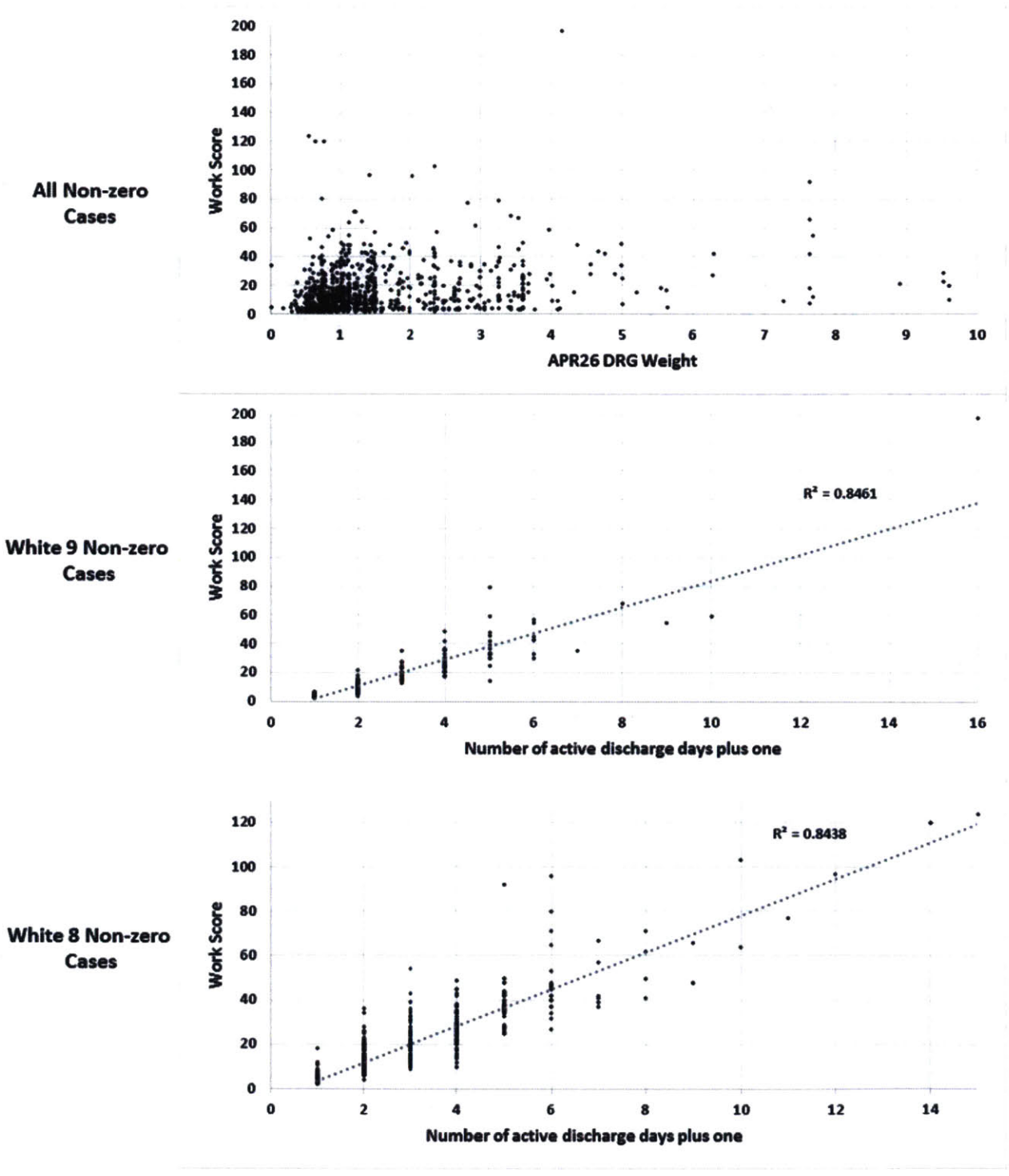


Figure B-1: Scatterplots of work scored versus CMI component patient DRG weights and a simple "Plus one" measure

Appendix C

Word Dictionary and Sub-Dictionaries Used for Analysis of Case Manager Note Text

The dictionary and sub-dictionaries used in the construction of the text feature vector for each of the case manager notes examined is provided in Table C.1. A count of each of the different word categories (e.g., cdword) formed one field of the feature vector. The count of all dictionary words provided another field of the feature vector. This aggregate word count is essentially the relevant length of a note from the perspective of automatically scoring workload. Chapter 4 provides a complete description of the text feature vector used in automatically scoring cases, retrospectively, based on the text of case manager notes.

Table C.1: Bag-of-words (BOW) dictionary and sub-disctionaries

Stand-alone words
met / meeting
spoke / spoken / speak
refer / referred / referral / referrals

Secondary words
call / called / calls / calling
phoned
fax / faxed
working
find / finding
investigating
check / checked / checking
addition / additional
meet

Other words
form
paperwork
submit / submitted
letter
documentation
pfs
hold
notes
message
email
apply / applied / application
complete / completed
order
voucher

Home words
delivery / delivered / deliver
order / ordered / orders

Transportation Words
schedule / scheduled
reschedule / rescheduled
taxi
cab
bus
ambulance
car
van
ride
arrange / arranged
book / booked
transport / transported / transportation

Consent words
agree / agrees / agreed / agreement / agreeable
amenable
consent / consents / consenting
aware
accept / accepts / acceptable
willing / willingness
acknowledges

Preference words
want / wants / wanting / wanted
wish / wishes
hope / hopes
desire / desires
prefer / prefers / preference
interest / interested
like / likes
choice / choices / choose / chose

Repeated words
need / needs / needed
require / requires / requiring / requirement
provide / provided

Communication / duration words
inform / informs / informed / informing
confirm / confirms / confirmed / confirming
tell / tells / telling / told
say / says / said /saying
discuss / discussed / discussing / discussion
inquire / inquired / inquiring
per
report / reports / reported / reporting / reportedly
conversation / conversations
encourage / encouraged
state / stated / states
clarify / clarified / clarifies / clarification
communicated
relate / relates / relating / related
relay / relays / related
shared
wonder / wonders / wondering
voice / voiced / voicing
verbalized / verbalized
reveal / reveals / revealed
rectify / rectified
advise / advised / advisement
obtain / obtained
reiterate / reiterated

Sentiment words type 1
refuse / refuses / refusing
resistant
reluctant
insist / insists / insisting / insisted
disagree / disagrees / disagreement
object / objects / objected / objecting
discouraged
unwilling / unwillingness

Trigger words
cancel / cancelled / cancelling
decide / decided / decision
elect / elected / electing / elects
offer / offers / offered / offering
hear / hearing / heard
reject / rejected / rejections
consider / considering / consideration
reconsider / reconsidered / reconsideration
talk / talking / talked
tour / toured / touring
request / requests / requesting / requested
recommend / recommends / recommended
give / gave / given
accept / accepted
deny / denied / denial / denials
decline / declined

Sentiment words type 2
concern / concerns / concerned
complain / complains / complaining / complaint
teary
worry / worried / worries
fear / fears / fearful
afraid
stress / stressed
overwhelmed

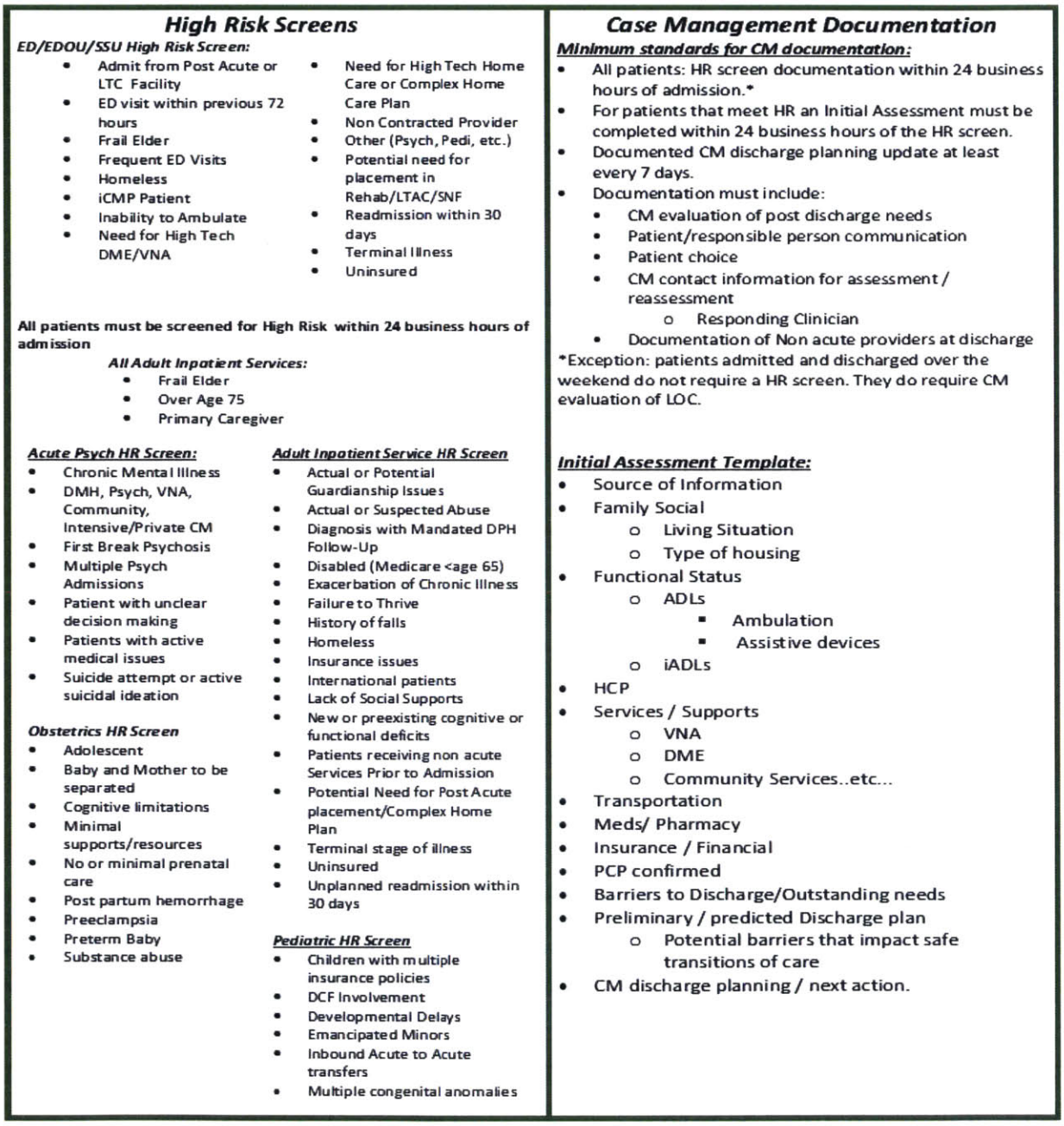
Sentiment words type 3
upset
frustrated
accuse / accuses
adamant / adamantly
stubborn
angry
dissatisfied / dissatisfaction

Appendix D

Case Management High Risk Screening Criteria and HRIA2 Initial Assessment Template

The criteria used to screen for high-risk patients are provided in Figure D-1. This screen is required within 24 business hours of patient admission. Like acuity, high-risk is a multi-dimensional concept. High-risk patients could be high-cost and/or high-volume patients[69]. In the context of our work, and in line with the use of the high-risk screen at MGH, high-risk refers to patients that may require case manager intervention.

Table D.1 shows one template used for the initial assessment, the HRIA2. The initial assessment is completed by case managers for patients, meeting high-risk criteria, within 48 business hours of admission. Even when the HRIA2 is used instead of, for example, a free-text narrative assessment, only a subset of questions typically have answers for a patient.



High Risk Screens

ED/EDOU/SSU High Risk Screen:

- Admit from Post Acute or LTC Facility
- ED visit within previous 72 hours
- Frail Elder
- Frequent ED Visits
- Homeless
- iCMP Patient
- Inability to Ambulate
- Need for High Tech DME/VNA
- Need for High Tech Home Care or Complex Home Care Plan
- Non Contracted Provider
- Other (Psych, Pedi, etc.)
- Potential need for placement in Rehab/LTAC/SNF
- Readmission within 30 days
- Terminal Illness
- Uninsured

All patients must be screened for High Risk within 24 business hours of admission

All Adult Inpatient Services:

- Frail Elder
- Over Age 75
- Primary Caregiver

Acute Psych HR Screen:

- Chronic Mental Illness
- DMH, Psych, VNA, Community, Intensive/Private CM
- First Break Psychosis
- Multiple Psych Admissions
- Patient with unclear decision making
- Patients with active medical issues
- Suicide attempt or active suicidal ideation

Obstetrics HR Screen

- Adolescent
- Baby and Mother to be separated
- Cognitive limitations
- Minimal supports/resources
- No or minimal prenatal care
- Post partum hemorrhage
- Preeclampsia
- Preterm Baby
- Substance abuse

Adult Inpatient Service HR Screen

- Actual or Potential Guardianship Issues
- Actual or Suspected Abuse
- Diagnosis with Mandated DPH Follow-Up
- Disabled (Medicare <age 65)
- Exacerbation of Chronic Illness
- Failure to Thrive
- History of falls
- Homeless
- Insurance issues
- International patients
- Lack of Social Supports
- New or preexisting cognitive or functional deficits
- Patients receiving non acute Services Prior to Admission
- Potential Need for Post Acute placement/Complex Home Plan
- Terminal stage of illness
- Uninsured
- Unplanned readmission within 30 days

Pediatric HR Screen

- Children with multiple insurance policies
- DCF Involvement
- Developmental Delays
- Emancipated Minors
- Inbound Acute to Acute transfers
- Multiple congenital anomalies

Case Management Documentation

Minimum standards for CM documentation:

- All patients: HR screen documentation within 24 business hours of admission.*
- For patients that meet HR an Initial Assessment must be completed within 24 business hours of the HR screen.
- Documented CM discharge planning update at least every 7 days.
- Documentation must include:
  - CM evaluation of post discharge needs
  - Patient/responsible person communication
  - Patient choice
  - CM contact information for assessment / reassessment
    - Responding Clinician
  - Documentation of Non acute providers at discharge

*Exception: patients admitted and discharged over the weekend do not require a HR screen. They do require CM evaluation of LOC.

Initial Assessment Template:

- Source of Information
- Family Social
  - Living Situation
  - Type of housing
- Functional Status
  - ADLs
    - Ambulation
    - Assistive devices
  - iADLs
- HCP
- Services / Supports
  - VNA
  - DME
  - Community Services..etc...
- Transportation
- Meds/ Pharmacy
- Insurance / Financial
- PCP confirmed
- Barriers to Discharge/Outstanding needs
- Preliminary / predicted Discharge plan
  - Potential barriers that impact safe transitions of care
- CM discharge planning / next action.

Figure D-1: High risk screening criteria

Table D.1: HRIA2 initial assessment template

SEQ	NO.	QUESTION	ANSWER CD	ANSWER
10	1	Reason unable to complete HR screen w/in 24 business hrs:	NOCHART	Chart Not Available
10	1	Reason unable to complete HR screen w/in 24 business hrs:	FSS	Fri, Sat, Sun-3 day Weekend Only
10	1	Reason unable to complete HR screen w/in 24 business hrs:	PTDISCH	Patient Discharged prior to CM Review
10	1	Reason unable to complete HR screen w/in 24 business hrs:	TRANS	Transfer from Another Unit
10	1	Reason unable to complete HR screen w/in 24 business hrs:	WORKLOAD	WorkLoad Issue
20	--	----------ED HIGH RISK SCREEN----------		
30	2	ED Patient identification:	CONSULT	Consult
30	2	ED Patient identification:	EDIS	EDIS
30	2	ED Patient identification:	ROUNDING	Rounding
30	2	ED Patient identification:	PRIORITY	Worklist/Priority Census
40	3	ED High Risk Screen (Select ALL that apply):	75YRS	75 Years Old –INACTIVATED
40	3	ED High Risk Screen (Select ALL that apply):	ADMITPA	Admit from Post-Acute or LTC Facility
40	3	ED High Risk Screen (Select ALL that apply):	ED72	ED Visit within Previous 72 Hrs
40	3	ED High Risk Screen (Select ALL that apply):	FRAIL	Frail Elder
40	3	ED High Risk Screen (Select ALL that apply):	FREQED	Frequent ED Visits
40	3	ED High Risk Screen (Select ALL that apply):	HOMELESS	Homeless
40	3	ED High Risk Screen (Select ALL that apply):	AMBUL	Inability to Ambulate
40	3	ED High Risk Screen (Select ALL that apply):	NHT	Need for High Tech DME/VNA
40	3	ED High Risk Screen (Select ALL that apply):	HTCHCP	Need for High Tech Home Care or Complex Home Care Plan
40	3	ED High Risk Screen (Select ALL that apply):	NCP	Non Contracted Provider
40	3	ED High Risk Screen (Select ALL that apply):	NOTHR	Not High Risk – NONE – INACTIVATED
40	3	ED High Risk Screen (Select ALL that apply):	OTHER	Other (Psych, Pedi, etc)
40	3	ED High Risk Screen (Select ALL that apply):	PNP	Potential Need for Placement in Rehab/LTAC/SNF
40	3	ED High Risk Screen (Select ALL that apply):	READMIT30	Readmission within 30 Days
40	3	ED High Risk Screen (Select ALL that apply):	TERMINAL	Terminal Illness
40	3	ED High Risk Screen (Select ALL that apply):	NOINSUR	Uninsured
40	3	ED High Risk Screen (Select ALL that apply):	CMPCMA	iCMP Patient
50	4	ED Interventions (Select ALL that apply):	ADMDIV	Admission Diversion
50	4	ED Interventions (Select ALL that apply):	COMPLEX	Complex Care Planning
50	4	ED Interventions (Select ALL that apply):	IPA	Initial Inpatient Assessment
50	4	ED Interventions (Select ALL that apply):	LOC	LOC Assessment
50	4	ED Interventions (Select ALL that apply):	FACILMED	Medications/Prescriptions

50	4	ED Interventions (Select ALL that apply):	NHT	Need for High Tech DME/VNA
50	4	ED Interventions (Select ALL that apply):	NOINT	No Intervention Required
50	4	ED Interventions (Select ALL that apply):	NONE	None
50	4	ED Interventions (Select ALL that apply):	NI	Nursing Intervention (Counsel, Education, etc)
50	4	ED Interventions (Select ALL that apply):	REFER	Refer for Consult-Other
50	4	ED Interventions (Select ALL that apply):	REFERPFS	Refer for Consult-PFS
50	4	ED Interventions (Select ALL that apply):	REFERSS	Refer for Consult-Social Services
50	4	ED Interventions (Select ALL that apply):	TRANS	Transportation Issues
60	5	ED Discharge ADA completed for patient discharged from the ED?		
70	6	ED High Risk Screen completed by:		
80	7	Date ED High Risk Screen completed:		
90	8	Patient meets ED High Risk Screen at ED Admission? (REQUIRED for ED patients)		
100	–	----------ACUTE CM HIGH RISK SCREEN----------		
170	15	Acute CM High Risk Screen completed by:		
180	16	Date Acute CM High Risk Screen completed:		
190	17	Patient meets Acute CM High Risk Screen at Admission? (REQUIRED for Acute Patients)		
200	18	Patient readmitted within last 6 months, the IA completed at that time was reviewed. Will you document with the Readmit Template?		
210	–	---------- INITIAL ASSESSMENT ----------		
215	19	Preadmission Comment:		
220	20	Information obtained from:	CMCC	Case Management/Care Coordination
220	20	Information obtained from:	CA	Community Agency
220	20	Information obtained from:	FA	Family
220	20	Information obtained from:	FR	Friend
220	20	Information obtained from:	LG	Legal Guardian
220	20	Information obtained from:	MD	MD
220	20	Information obtained from:	MR	Medical Record
220	20	Information obtained from:	P	Patient
220	20	Information obtained from:	RN	RN
220	20	Information obtained from:	SO	Significant Other
220	20	Information obtained from:	S	Spouse
230	21	Please enter Health Care Proxy name if known:		
240	22	Please enter Guardian's name if known:		
250	23	Can the patient read and write -Please describe if applicable:		

260	24	Primary language, interpreter needed/used - Please describe if applicable:		
270	25	If admitted from another acute/non-acute facility, please specify facility name:		
280	26	SNF/NF Comments (ie bed held, private pay, etc):		
290	–	----------Living Arrangements----------		
300	27	Permanent living situation:	ALF	Assisted Living Facility
300	27	Permanent living situation:	EDH	Elderly/Disabled Housing
300	27	Permanent living situation:	FH	Foster Home
300	27	Permanent living situation:	GH	Group Home
300	27	Permanent living situation:	H	Homeless
300	27	Permanent living situation:	I	Incarcerated
300	27	Permanent living situation:	IH	Independent Housing
300	27	Permanent living situation:	NH	Nursing Home
300	27	Permanent living situation:	RH	Rooming House
300	27	Permanent living situation:	S	Shelter
300	27	Permanent living situation:	UH	Unstable Housing
310	28	# of stairs to enter into home:		
320	29	# of stairs inside home:		
330	30	Any devices to help negotiate stairs:	CL	Chair Lift
330	30	Any devices to help negotiate stairs:	E	Elevator
330	30	Any devices to help negotiate stairs:	HR	Hand Rails
330	30	Any devices to help negotiate stairs:	R	Ramp
340	31	Bedroom(s) location:	1	1st Floor
340	31	Bedroom(s) location:	2	2nd Floor or Higher
340	31	Bedroom(s) location:	B	Both
350	32	Bathroom(s) location:	1	1st Floor
350	32	Bathroom(s) location:	2	2nd Floor or Higher
350	32	Bathroom(s) location:	B	Both
360	33	Home setup comment (Bed/bath location, DME, etc):		
370	34	With whom does patient live with:	AC	Adult Child(ren)
370	34	With whom does patient live with:	NO	Alone
370	34	With whom does patient live with:	FP	Foster Parent(s)
370	34	With whom does patient live with:	F	Friend(s)
370	34	With whom does patient live with:	GP	Grandparent(s)
370	34	With whom does patient live with:	MC	Minor Child(ren)
370	34	With whom does patient live with:	OR	Other Residents

370	34	With whom does patient live with:	P	Parent(s)
370	34	With whom does patient live with:	PSO	Partner/Significant Other
370	34	With whom does patient live with:	PE	Pet(s)
370	34	With whom does patient live with:	PC	Private Caretaker
370	34	With whom does patient live with:	R	Roommate(s)
370	34	With whom does patient live with:	SI	Sibling(s)
370	34	With whom does patient live with:	S	Spouse
380	35	If anyone depends on patient for care, please describe:		
390	36	Describe patient's work/education status (disabled, full/part time, retired, student etc):		
400	--	----------Functional Assessment----------		
410	37	Indoor mobility at baseline:	BB	Bed Bound
410	37	Indoor mobility at baseline:	C	Cane
410	37	Indoor mobility at baseline:	CR	Crutches
410	37	Indoor mobility at baseline:	FW	Furniture Walking
410	37	Indoor mobility at baseline:	I	Independent
410	37	Indoor mobility at baseline:	MS	Motorized Scooter
410	37	Indoor mobility at baseline:	PCA	Primary Caretaker Assisted (Pedi)
410	37	Indoor mobility at baseline:	W	Walker
410	37	Indoor mobility at baseline:	WB	Wheelchair Bound
420	38	Outdoor mobility at baseline	BB	Bed Bound
420	38	Outdoor mobility at baseline	C	Cane
420	38	Outdoor mobility at baseline	CR	Crutches
420	38	Outdoor mobility at baseline	I	Independent
420	38	Outdoor mobility at baseline	MS	Motorized Scooter
420	38	Outdoor mobility at baseline	NA	Non Ambulatory
420	38	Outdoor mobility at baseline	PCA	Primary Caretaker Assisted (Pedi)
420	38	Outdoor mobility at baseline	W	Walker
420	38	Outdoor mobility at baseline	WCB	Wheel Chair Bound
425	38b	Mobility Comment:		
430	39	Is patient independent with ADLs?		
440	40	Feeding:	FD	Fully Dependent
440	40	Feeding:	GT	G-Tube
440	40	Feeding:	I	Independent
440	40	Feeding:	JT	J-Tube
440	40	Feeding:	NA	Needs Assistance
440	40	Feeding:	SF	Special Formulas (PO)

440	40	Feeding:	TPN	TPN
450	41	Toileting:	FD	Fully Dependent
450	41	Toileting:	I	Independent
450	41	Toileting:	NA	Needs Assistance
460	42	Bathing/grooming:	FD	Fully Dependent
460	42	Bathing/grooming:	I	Independent
460	42	Bathing/grooming:	NA	Needs Assistance
470	43	Dressing:	FD	Fully Dependent
470	43	Dressing:	I	Independent
470	43	Dressing:	NA	Needs Assistance
480	44	Comment on ADLs above:		
490	45	Patient is independent with IADLs?		
500	46	Financial management (pay bills)?	DNP	Does Not Perform
500	46	Financial management (pay bills)?	I	Independent
500	46	Financial management (pay bills)?	NA	Needs Assistance
510	47	Shopping:	DNP	Does Not Perform
510	47	Shopping:	I	Independent
510	47	Shopping:	NA	Needs Assistance
520	48	House keeping/laundry:	DNP	Does Not Perform
520	48	House keeping/laundry:	I	Independent
520	48	House keeping/laundry:	NA	Needs Assistance
530	49	Meal preparation:	DNP	Does Not Perform
530	49	Meal preparation:	I	Independent
530	49	Meal preparation:	NA	Needs Assistance
540	50	Medication administration:	DNP	Does Not Perform
540	50	Medication administration:	I	Independent
540	50	Medication administration:	NA	Needs Assistance
550	51	Comment on IADLs above:		
560	–	----------Continuum of Care----------		
570	52	Who will assist patient at discharge:		
580	53	Usual mode of transportation:	A	Ambulance
580	53	Usual mode of transportation:	D	Disability (The Ride)
580	53	Usual mode of transportation:	F	Family/Friend
580	53	Usual mode of transportation:	M	Medicaid (e.g. PT-1)
580	53	Usual mode of transportation:	PD	Public Transportation (Bus, Train, T, etc)
580	53	Usual mode of transportation:	S	Self

580	53	Usual mode of transportation:	TC	Taxi/Cab
590	54	Does patient have transportation to home at discharge & for follow up appts? Please describe:		
600	55	Does patient have PCP and is name correct?		
610	56	Please describe plan to address incorrect/no PCP:		
620	57	Is the health insurance correct and adequate to cover potential DC needs?		
630	58	Health insurance comment/Specialty Case Manager, if known:		
640	59	Prescription coverage:	HI	Health Insurance
640	59	Prescription coverage:	HSN	Health Safety Net
640	59	Prescription coverage:	MPD	Medicare Part D
640	59	Prescription coverage:	N	None
640	59	Prescription coverage:	PA	Prescription Advantage
640	59	Prescription coverage:	U	Unknown
640	59	Prescription coverage:	VA	Veterans Administration
650	60	Prescription comment (i.e Difficulty Obtaining Meds, etc)		
660	61	Name, location and phone number(s) of pharmacies used:		
670	62	ACTIVE in-home services and agencies list if known (ie VNA, DME, Respiratory, Elder Services, Haven etc):		
680	63	INACTIVE home services, list if known:		
690	64	Inpatient SNF/REHAB/Non Acute facilities and type, list if known:		
700	65	Other OP Services/Community Agencies:	ADH	Adult Day Health
700	65	Other OP Services/Community Agencies:	C	Chemo
700	65	Other OP Services/Community Agencies:	CB	Commission for the Blind
700	65	Other OP Services/Community Agencies:	CDHH	Commission for the Deaf and Hard of Hearing
700	65	Other OP Services/Community Agencies:	CCM	Community Case Manager
700	65	Other OP Services/Community Agencies:	CM	Coumadin Management
700	65	Other OP Services/Community Agencies:	DDS	Dept of Developmental Services
700	65	Other OP Services/Community Agencies:	DMH	Dept of Mental Health
700	65	Other OP Services/Community Agencies:	D	Dialysis
700	65	Other OP Services/Community Agencies:	EI	Early Intervention
700	65	Other OP Services/Community Agencies:	MH	Mental Health
700	65	Other OP Services/Community Agencies:	NA	Not Applicable
700	65	Other OP Services/Community Agencies:	PTOT	PT/OT
700	65	Other OP Services/Community Agencies:	PCAS	Personal Care Attendant Services
700	65	Other OP Services/Community Agencies:	R	Radiation
710	66	Other OP Services/Community Agencies Additional Comments:		

720	–	----------Anticipated DC Plan----------		
730	67	Identified barriers/concerns related to discharge:	CNTP	Compliance with Needed Treatment Plan
730	67	Identified barriers/concerns related to discharge:	CPHDV	Current or Past Hx of Domestic Violence
730	67	Identified barriers/concerns related to discharge:	IS	Immigration Status that may Impact Access to DC Services
730	67	Identified barriers/concerns related to discharge:	IIC	Inadequate Insurance Coverage to Cover Likely DC Needs
730	67	Identified barriers/concerns related to discharge:	ISS	Inadequate Social Supports to Support Likely DC Needs
730	67	Identified barriers/concerns related to discharge:	PCI	Potential Capacity Issue with No Official Alternate Decision Maker
730	67	Identified barriers/concerns related to discharge:	SEN	Specialty Equipment Needs (i.e Bariatrics, etc)
740	68	Issues/problems - additional information:		
750	69	Will refer to:	A	Admitting
750	69	Will refer to:	C	Chaplaincy
750	69	Will refer to:	FC	Financial Counseling
750	69	Will refer to:	H	Haven
750	69	Will refer to:	N	Nutrition
750	69	Will refer to:	SW	Social Work
760	70	Explain reason for each referral above:		
770	71	Will request MD order for:	OT	OT
770	71	Will request MD order for:	PT	PT
770	71	Will request MD order for:	PC	Palliative Care
770	71	Will request MD order for:	PH	Physiatry
770	71	Will request MD order for:	PS	Psychiatry
770	71	Will request MD order for:	S	Speech
780	72	Preliminary discharge plan:	AH	Acute Hospital
780	72	Preliminary discharge plan:	BTH	Bridge to Hospice
780	72	Preliminary discharge plan:	H	Home
780	72	Preliminary discharge plan:	HWS	Home With Services
780	72	Preliminary discharge plan:	HOS	Hospice
780	72	Preliminary discharge plan:	IRF	Inpatient Rehab Facility
780	72	Preliminary discharge plan:	LTAC	Long Term Acute Care
780	72	Preliminary discharge plan:	NRE	Need to Re-Evaluate
780	72	Preliminary discharge plan:	OC	Outpatient Care
780	72	Preliminary discharge plan:	RTLTC	Return to Long Term Care
780	72	Preliminary discharge plan:	S	Shelter
780	72	Preliminary discharge plan:	SNF	Skilled Nursing Facility

780	72	Preliminary discharge plan:	UDTT	Unable to Determine at This Time
780	72	Preliminary discharge plan:	VM	Visiting Moms
780	72	Preliminary discharge plan:	WIC	Women, Infants and Children
790	73	**Summary Comment supporting current Level of Care and projected transitional goals:**		
800	74	Plan discussed with team:	W	Will Discuss with Team
800	74	Plan discussed with team:	Y	Yes - Discussed with Team
810	75	Patient/family agree with preliminary discharge plan?		
820	76	If patient/family does NOT agree with preliminary discharge plan, please describe:		
830	–	----------Psych Specialty Questions----------		
840	77	Name and phone number(s) of Psychiatrist/Therapist, if known:		
850	78	Please list PHP/IOP, if known:		
860	79	Please list inpatient detox, if known:		
870	–	----------Pedi/OB Specialty Questions----------		
880	80	Is a car seat available?		
890	81	Any history of compound meds?		
900	–	----------Attestation----------		
910	82	Assessment started by:		
920	83	Date started:		
930	84	Assessment completed by:		
940	85	Date completed:		
950	86	Edited by:		
960	87	Edited on:		
970	88	Misc Info: include Primary Contact (will print on Scut Census):		

Appendix E

Discharge planning and utilization review: Case manager positions and associated caseloads

The DCP case manager positions and number of assigned beds are shown in Table E.1; Table E.2 provides comparable information for UR case managers. This is a composite view, constructed by examining daily weekday schedules for the month of May 2015. During the course of our work 52-53 case managers staffed the numbered positions on a typical weekday. On weekends, 5-6 case managers provided coverage. The case manager positions also had one of seven case management resource specialists assigned for support as indicated in the tables. Color-coding is used to highlight case manager positions with assigned beds on more than one floor of MGH.

Table E.1: Discharge planning case manager positions at MGH

Building	Floor	Service	Case Manger Position	Beds	CMSU Support
Blake	6	Transplant	1	21	
Blake	7	Medical ICU	2	11	2
			3	7	
Blake	8	Cardiac surgery ICU	1	5	
			4	4	
			5	4	
			6	5	
Blake	10	Neonatal ICU	7	18	3
Blake	11	Psychiatry	8	24	4
Blake	12	ICU	9	18	4
Blake	13	Obstetrics / SCN	10	21 / 24	3
Blake	14	Labor/Delivery	10		3
Gray/Bigelow	6	Pediatric ICU	11	14	
Gray/Bigelow	7	Short stay unit	12	18	
Gray/Bigelow	9	Respiratory acute care (10) / Medicine (8)	13	18	
Gray/Bigelow	11	Medicine	14	25	2
Gray/Bigelow	12	ED observation unit	15	14	
Emergency Dept			16		
Gray/Bigelow	14	Vascular	17	27	7
Ellison	3	PACU	18		7
Ellison	4	Surgical ICU	9	14	
			18	6	
Ellison	6	Orthopedics (27) / Urology (9)	19	24	
			20	12	
Ellison	7	General surgery (34) / Urology(2)	18	12	7
			21	24	
Ellison	8	Cardiac surgery (30) / Intervention (6)	22	29	
			23	5	
Ellison	9	Medicine: Critical care unit	22	4	
			23	5	
			24	4	
Ellison	10	Medicine: Step-down unit	23	5	
			25	26/8	
Ellison	11	Medicine: Cardiac intervention	23	5	
			24	29	
Ellison	12	Medicine	2	6	
			26	30	

Building	Floor	Service	Case Manger Position	Beds	CMSU Support
Ellison	13	Obstetrics	10	14/6/16	3
Ellison	14	Burn unit / Plastics	6	21	3
Ellison	16	Medicine	2	6	
			27	30	
Ellison	17	Pediatrics	7	20	
Ellison	18	Pediatrics	11	24	
Ellison	19	Thoracic surgery (20) / Medicine (10)	28	30	7
Lunder	6	Neuro ICU	29	22	4
Lunder	7	Neuro	29	6	4
			30	18	
Lunder	8	Neuro	32	21	4
Lunder	9	Oncology	33	22	4
Lunder	10	Oncology	35	22	4
Phillips	20	Medicine	3	26	
Phillips	21	Gynecology	4	20	7
Phillips	22	Surgery (10) /Medicine (7) / Ortho (2)	5	19	7
White	6	Ortho (28) / OMF (2)	20	9	
			36	21	
White	7	General surgery	18	3	7
			37	24	
White	8	Medicine	38	24	2
			39	2	
White	9	Medicine	40	25	2
White	10	Medicine	39	20	2
White	11	Medicine	41	24	2

A	SNF waiver
B	Late day bed offer
C	Ortho/RAPT
D	Psych consult

Table E.2: Utilization review case manager positions at MGH

Building	Floor	Service	Case Manger Position	Beds	CMSU Support
Blake	6	Transplant	1	21	1
Blake	7	Medical ICU	2	18	2
Blake	8	Cardiac surgery ICU	2	18	1
Blake	10	Neonatal ICU	3	18	3
Blake	11	Psychiatry	8	24	4
Blake	12	ICU	4	18	4
Blake	13	Obstetrics / SCN	10	21 / 24	3
Blake	14	Labor/Delivery	10		3
Gray/Bigelow	6	Pediatric ICU	3	14	5
Gray/Bigelow	9	Respiratory acute care (10) / Medicine (8)	5	18	1
Gray/Bigelow	11	Medicine	6	25	2
Gray/Bigelow	14	Vascular	7	27	7
Ellison	4	Surgical ICU	9	20	1
Ellison	6	Orthopedics (27) / Urology (9)	11	36	3
Ellison	7	General surgery (34) / Urology(2)	4	36	7
Ellison	8	Cardiac surgery (30) / Intervention (6)	12	36	1
Ellison	9	Medicine: Critical care unit	13	16	1
Ellison	10	Medicine: Step-down unit	13	36	1
Ellison	11	Medicine: Cardiac intervention	13	36	1
Ellison	12	Medicine	13	36	1

Building	Floor	Service	Case Manger Position
Ellison	13	Obstetrics	10
Ellison	14	Burn unit / Plastics	15
Ellison	16	Medicine	14
Ellison	17	Pediatrics	3
Ellison	18	Pediatrics	3
Ellison	19	Thoracic surgery (10) / Medicine (10)	7
Lunder	6	Neuro ICU	5
Lunder	7	Neuro	16
Lunder	8	Neuro	16
Lunder	9	Oncology	17
Lunder	10	Oncology	17
Phillips	20	Medicine	14
Phillips	21	Gynecology	18
Phillips	22	Surgery (10) /Medicine (7) / Ortho (2)	19
White	6	Ortho (28) / OMF (2)	11
White	7	General surgery	9
White	8	Medicine	2
White	9	Medicine	5
White	10	Medicine	2
White	11	Medicine	6

Bibliography

[1] Alecia Laing, RN, Case Manager at MGH. Interview conducted by Jason Stuck on 20150714.

[2] Angel Figueroa, CMRS Manager at MGH. Interview conducted by Jason Stuck on 20151015.

[3] Lecture on "Electrical Units of Measurement" (3 May 1883). published in Popular Lectures Vol. I, p. 73.

[4] Michael Trotta, RN, Case Manager at MGH. Interview conducted by Jason Stuck on 20150630.

[5] MIT-MGH Collaboration Case Management Meetings. Meetings held during Fall of 2015.

[6] Nancy Sullivan, Executive Director of Case Management at MGH; Rachael McKenzie, RN, Nurse Director of Case Management at MGH; Debra Connolly, RN, Nurse Director of Case Management at MGH. Interview conducted by Jason Stuck on 20150706.

[7] Nora Arbeene, RN, Case Manager at MGH. Interviews conducted by Jason Stuck on 20150629 and 20151107.

[8] Rachael McKenzie, RN, Nurse Director of Case Management at MGH. Interview conducted by Jason Stuck on 20150629.

[9] Sheila Cambra, RN, Case Manager Clinical Specialist at MGH. Interview conducted by Jason Stuck on 20150827.

[10] 42 CFR 482.43 - CONDITION OF PARTICIPATION: DISCHARGE PLANNING. `https://www.gpo.gov/fdsys/granule/CFR-2011-title42.html`, 1994, ammended 2004. [Online; accessed 20160511].

[11] *Standards of Practice for Case Management.* Case Management Society of America, 2002.

[12] *Assign cases based on acuity level of patients, not just on the number*, volume 11. Hospital Case Management, 2003.

[13] *Avoid overload: Assign cases based on workload, model, and role functions*, volume 20. Hospital Case Management, 2005.

[14] *Case Management Caseload Concept Paper: Proceedings of the Caseload Work Group, a Joint Collaboration of CMSA and NASW.* Case Management Society of America, 2008.

[15] *Standards of Practice for Case Management.* Case Management Society of America, 2010.

[16] What the Heck is a DRG? And Why Should I Care About Case Mix? http://codercoach.blogspot.com/2011/01/what-heck-is-drg-and-why-should-i-care.html, 2011. [Online; accessed 20160731].

[17] Definition of Case Mix for hospitals. http://www.healthandhospitalcommission.com/docs/May26Meeting/CasemixIndexDefintion.pdf, 2012. [Online; accessed 20160731].

[18] *Inpatient vs. observation: Get it right the first time*, volume 20. Hospital Case Management, 2012.

[19] Medical necessity & charting guidelines. https://www.utcomchatt.org/docs/Medical_Necessity_and_charting_guidelines.pdf, 2013. [Online; accessed 20160605].

[20] A Patient's Guide to Observation Care. http://www.ihatoday.org/uploadDocs/1/observationstayguidelines.pdf, 2014. [Online; accessed 2016049].

[21] American Academy of Family Physicians – Risk-stratified care management. https://www.anthem.com/provider/noapplication/f1/s0/t0/pw_e225424.pdf?refer=ehpprovider, 2014. [Online; accessed 20160505].

[22] Are you an inpatient or an outpatient? https://www.medicare.gov/Pubs/pdf/11435.pdf, 2014. [Online; accessed 20160506].

[23] Analyzing case mix index and the impact on CDI programs. https://www.huronconsultinggroup.com/Insights/Whitepapers/Healthcare/~/media/47777A4ABB824A3B9A9E7FBFF51E4DE9.ashx, 2015. [Online; accessed 20160731].

[24] Clinical and Practice Management:Utilization Review FAQ. https://www.acep.org/Clinical---Practice-Management/Utilization-Review-FAQ, 2015. [Online; accessed 20160501].

[25] What Utilization Management Is and How Decisions Are Made. http://healthamerica.coventryhealthcare.com/web/groups/public/@cvty_regional_chcpa/documents/document/c041883.pdf, 2015. [Online; accessed 20160502].

[26] Example of Usage: The Case Mix Index-Office of Statewide Health, Planning, and, Development. http://www.oshpd.ca.gov/HID/Products/PatDischargeData/CaseMixIndex/CMI/ExampleCalculation.pdf, 2016. [Online; accessed 20160501].

[27] Hospital Overview: Massachusetts General Hospital. http://www.massgeneral.org/about/overview.aspx, 2016. [Online; accessed 20160501].

[28] Medicare's Skilled Nursing Facility (SNF) Three-Day Inpatient Stay Requirement: In Brief. https://www.ahcancal.org/advocacy/solutions/Documents/Congressional/Requirement.pdf, 2016. [Online; accessed 20160608].

[29] MIT LGO Partner Companies. http://lgo.mit.edu/partner-companies/partners, 2016. [Online; accessed 20160312].

[30] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*, pages 30–38. Association for Computational Linguistics, 2011.

[31] Alan Agresti and Maria Kateri. *Categorical data analysis.* Springer, 2011.

[32] Linda H Aiken, Sean P Clarke, Douglas M Sloane, Julie Sochalski, and Jeffrey H Silber. Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *Jama*, 288(16):1987–1993, 2002.

[33] Elizabeth Armstrong, Monique C de Waard, Harm-Jan S de Grooth, Martijn W Heymans, Dinis Reis Miranda, Armand RJ Girbes, and Jan Jaap Spijkstra. Using nursing activities score to assess nursing workload on a medium care unit. *Anesthesia & Analgesia*, 121(5):1274–1280, 2015.

[34] David H Autor, Frank Levy, and Richard J Murnane. The skill content of recent technological change: An empirical exploration. Technical report, National Bureau of Economic Research, 2001.

[35] David H Autor and Brendan Price. The changing task composition of the us labor market: An update of autor, levy, and murnane (2003). *Unpublished manuscript*, 2013.

[36] RF Avrill, N Goldfield, JS Huges, et al. All patient refined diagnosis related groups (apr-drgs) version 20.0: Methodology overview. *Wallingford, CT: 3M Health Information Systems*, 2003.

[37] Sarah F Baillon, Rosemary G Simpson, Nicky J Poole, Rebecca J Colledge, Nick A Taub, and Richard J Prettyman. The development of a scale to aid caseload weighting in a community mental health team for older people. *Journal of Mental Health*, 18(3):253–261, 2009.

[38] Amy Balstad and Pam Springer. Quantifying case management workloads: Development of the pace tool. *Professional Case Management*, 11(6):291–302, 2006.

[39] Cosmin A Bejan, Lucy Vanderwende, Heather L Evans, Mark M Wurfel, and Meliha Yetisgen-Yildiz. On-time clinical phenotype prediction based on narrative reports. In *AMIA Annual Symposium Proceedings*, volume 2013, page 103. American Medical Informatics Association, 2013.

[40] Michael W Berry and Malu Castellanos. Survey of text mining. *Computing Reviews*, 45(9):548, 2004.

[41] J Martin Bland and Douglas G Altman. The odds ratio. *Bmj*, 320(7247):1468, 2000.

[42] ANNE BRADY, Gobnait Byrne, Paul Horan, Colin Griffiths, Catriona Macgregor, and Cecily Begley. Measuring the workload of community nurses in ireland: a review of workload measurement systems. *Journal of nursing management*, 15(5):481–489, 2007.

[43] Paula Branco, Luis Torgo, and Rita Ribeiro. A survey of predictive modelling under imbalanced distributions. *arXiv preprint arXiv:1505.01658*, 2015.

[44] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees.* CRC press, 1984.

[45] Caitlin W Brennan and Barbara J Daly. Patient acuity: A concept analysis. *Journal of advanced nursing*, 65(5):1114–1126, 2009.

[46] Ana Paula de Brito and Edinêis de Brito Guirardello. Nursing workload in an inpatient unit. *Revista latino-americana de enfermagem*, 19(5):1139–1145, 2011.

[47] T Campbell, S Taylor, S Callaghan, and C Shuldham. Case mix type as a predictor of nursing workload. *Journal of Nursing Management*, 5(4):237–240, 1997.

[48] Francisco Javier Carmona-Monge, Gloria M^{a} Rollán Rodríguez, Cristina Quirós Herranz, Sonia García Gómez, and Dolores Marín-Morales. Evaluation of the nursing workload through the nine equivalents for nursing manpower use scale and the nursing activities score: a prospective correlation study. *Intensive and Critical Care Nursing*, 29(4):228–233, 2013.

[49] T Cesta. Case management insider. a further look into case management roles, functions, models, and case loads. *Hospital case management: the monthly update on hospital-based care planning and critical paths*, 20(2):23–24, 2012.

[50] T Cesta. The role of case management in an era of healthcare reform–part 1. *Hospital case management: the monthly update on hospital-based care planning and critical paths*, 20(7):103–106, 2012.

[51] T Cesta. The role of case management in an era of healthcare reform–part 2. *Hospital case management: the monthly update on hospital-based care planning and critical paths*, 20(8):119–122, 2012.

[52] T Cesta. The role of case management in an era of healthcare reform–part 3. *Hospital case management: the monthly update on hospital-based care planning and critical paths*, 20(9):135–138, 2012.

[53] N Chawla. Mining when classes are imbalanced, rare events matter more, and errors have costs attached. In *Proc. of SDM*, 2009.

[54] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[55] Nitesh V Chawla, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 107–119. Springer, 2003.

[56] Benjamin Arthur Christensen. Improving icu patient flow through discrete-event simulation. Master's thesis, Massachusetts Institute of Technology, 2012.

[57] David A Cieslak, T Ryan Hoens, Nitesh V Chawla, and W Philip Kegelmeyer. Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery*, 24(1):136–158, 2012.

[58] Brian Connolly, Pawel Matykiewicz, K Bretonnel Cohen, Shannon M Standridge, Tracy A Glauser, Dennis J Dlugos, Susan Koh, Eric Tham, and John Pestian. Assessing the similarity of surface linguistic features related to epilepsy across pediatric hospitals. *Journal of the American Medical Informatics Association*, 21(5):866–870, 2014.

[59] Kathy Craig and Anne Flaherty-Quemere. Implementing an automated acuity tool for scoring case management cases and caseloads at blue cross blue shield of massachusetts. *Professional case management*, 14(4):185–191, 2009.

[60] Kathy Craig and Diane L Huber. Acuity and case management: A healthy dose of outcomes, Part II. *Professional case management*, 12(4):199–210, 2007.

[61] Licong Cui, Samden D Lhatoo, Guo-Qiang Zhang, Satya Sanket Sahoo, and Alireza Bozorgi. Epidea: extracting structured epilepsy and seizure information from patient discharge summaries for cohort identification. In *AMIA*, 2012.

[62] Andrea Dal Pozzolo, Olivier Caelen, Serge Waterschoot, and Gianluca Bontempi. Racing for unbalanced methods selection. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 24–31. Springer, 2013.

[63] RW Day and GP Quinn. Comparisons of treatments after an analysis of variance in ecology. *Ecological monographs*, 59(4):433–463, 1989.

[64] Donna Diers and Janis Bozzo. Nursing resource definition in drgs. *Nursing Economics*, 15(3):124–132, 1997.

[65] Sara A Dolcetti. Analyzing the impact of delays for patient transfers from the icu to general care units. Master's thesis, Massachusetts Institute of Technology, 2015.

[66] Sandra R Edwardson and Phyllis B Giovannetti. Nursing workload measurement systems. *Annual review of nursing research*, 12:95, 1994.

[67] Wei Fan, Salvatore J Stolfo, Junxin Zhang, and Philip K Chan. Adacost: misclassification cost-sensitive boosting. In *Icml*, pages 97–105, 1999.

[68] Ingo Feinerer. Introduction to the tm package text mining in r, 2015.

[69] Anita Ward Finkelman. *Case management for nurses.* Pearson, 2011.

[70] Centers for Disease Control, Prevention, et al. National hospital discharge survey: 2010. *Atlanta (GA): CDC [online]. Available from URL: http://www. cdc. gov/nchs/nhds. htm.[Accessed 20160502]*, 2014.

[71] George Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, 3(Mar):1289–1305, 2003.

[72] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *Icml*, volume 96, pages 148–156, 1996.

[73] Milton Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.

[74] Daniela C Gonçalves-Bradley, Natasha A Lannin, Lindy M Clemson, Ian D Cameron, and Sasha Shepperd. Discharge planning from hospital. *The Cochrane Library*, 2016.

[75] Margaret Jean Hall, Carol J DeFrances, Sonja N Williams, Aleksandr Golosinskiy, Alexander Schwartzman, et al. National hospital discharge survey: 2007 summary. *Natl Health Stat Report*, 29(29):1–20, 2010.

[76] Viktor J Hansen, Kirill Gromov, Lauren M Lebrun, Harry E Rubash, Henrik Malchau, and Andrew A Freiberg. Does the risk assessment and prediction tool predict discharge disposition after joint replacement? *Clinical Orthopaedics and Related Research®*, 473(2):597–601, 2015.

[77] P. E. Hart. The condensed nearest neighbor rule. *IEEE transactions on information theory*, 1968.

[78] Haibo He and Yunqian Ma. *Imbalanced learning: foundations, algorithms, and applications.* John Wiley & Sons, 2013.

[79] Jonas Hiltrop. Modeling neuroscience patient flow and inpatient bed management. Master's thesis, Massachusetts Institute of Technology, 2014.

[80] T Ryan Hoens, NITESH V Chawla, H He, and Y Ma. Imbalanced datasets: from sampling to classifiers. *Imbalanced Learning: Foundations, Algorithms and Applications. Wiley*, pages 43–59, 2013.

[81] Diane L Huber and Kathy Craig. Acuity and case management: A healthy dose of outcomes, Part I. *Professional case management*, 12(3):132–144, 2007.

[82] Diane L Huber and Kathy Craig. Acuity and case management: A healthy dose of outcomes, Part III". *Professional case management*, 12(5):254–269, 2007.

[83] Matthew Hughes. Nursing workload: an unquantifiable entity. *Journal of Nursing Management*, 7(6):317–322, 1999.

[84] Beatrice Kalisch, Christopher R Friese, Seung Hee Choi, and Monica Rochman. Hospital nurse staffing: choice of measure matters. *Medical care*, 49(8):775, 2011.

[85] Maurice G Kendall and B Babington Smith. The problem of m rankings. *The annals of mathematical statistics*, 10(3):275–287, 1939.

[86] Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, 11:538–541, 2011.

[87] Miroslav Kubat, Stan Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *ICML*, volume 97, pages 179–186. Nashville, USA, 1997.

[88] Max Kuhn and Kjell Johnson. *Applied predictive modeling.* Springer, 2013.

[89] Joanie Lachance, Frédéric Douville, Clémence Dallaire, Katia Grillo Padilha, and Maria Cecilia Gallani. The use of the nursing activities score in clinical settings: an integrative review. *Revista da Escola de Enfermagem da USP*, 49(SPE):147–156, 2015.

[90] Corine Latour, Rebecca Perez, Roger G Kathol, Frits Huyse, Janice S Cohen, et al. *The integrated case management manual: Assisting complex patients regain physical and mental health.* Springer Publishing Company, 2010.

[91] Constance Lechman. The development of a caseload weighting tool. *Administration in Social Work*, 30(2):25–37, 2006.

[92] Jackson Liscombe, Giuseppe Riccardi, and Dilek Z Hakkani-Tür. Using context to improve emotion detection in spoken dialog systems. In *Interspeech*, pages 1845–1848, 2005.

[93] Jonas F Ludvigsson, Jyotishman Pathak, Sean Murphy, Matthew Durski, Phillip S Kirsch, Christophe G Chute, Euijung Ryu, and Joseph A Murray. Use of computerized algorithm to identify individuals in need of testing for celiac disease. *Journal of the American Medical Informatics Association*, 20(e2):e306–e310, 2013.

[94] Justin Martineau and Tim Finin. Delta tfidf: An improved feature space for sentiment analysis. *ICWSM*, 9:106, 2009.

[95] Barbara Mawn. Trends in case management acuity determination. *Occupational Medicine & Health Affairs 2013*, 2013.

[96] Mary Jane McKendry and Judy Van Horn. Today's hospital-based case manager: How one hospital integrated/adopted evidenced-based medicine using interqual® criteria. *Professional Case Management*, 9(2):61–71, 2004.

[97] David Meyer, Kurt Hornik, and Ingo Feinerer. Text mining infrastructure in r. *Journal of statistical software*, 25(5):1–54, 2008.

[98] Dinis Reis Miranda, Raoul Nap, Angelique de Rijk, Wilmar Schaufeli, Gaetano Iapichino, Members of the TISS Working Group, et al. Nursing activities score. *Critical care medicine*, 31(2):374–382, 2003.

[99] A Jacqueline Mitus. The birth of interqual: Evidence-based decision support criteria that helped change healthcare. *Professional case management*, 13(4):228–233, 2008.

[100] Roisin Morris, Padraig MacNeela, Anne Scott, Pearl Treacy, and Abbey Hyde. Reconsidering the conceptualization of nursing workload: literature review. *Journal of advanced Nursing*, 57(5):463–471, 2007.

[101] Catherine M Mullahy. Case management and managed care. *The Managed health care handbook*, pages 274–300, 1996.

[102] Tetsuya Nasukawa and Tohru Nagano. Text analysis and knowledge mining system. *IBM systems journal*, 40(4):967–984, 2001.

[103] L O'Brien-Pallas. An analysis of the multiple approaches to measuring nursing workload. *Canadian journal of nursing administration*, 1(2):8–11, 1988.

[104] Linda O'Brien-Pallas, Rhonda Cockerill, and Peggy Leatt. Different systems, different costs?: An examination of the comparability of workload measurement systems. *Journal of Nursing Administration*, 22(12):17–22, 1992.

[105] Linda O'Brien-Pallas, Diane Irvine, Elisabeth Peereboom, and Michael Murray. Measuring nursing workload: understanding the variability. *Nursing economics*, 15(4):171–182, 1997.

[106] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

[107] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

[108] Devon Jameson Price. Managing variability to improve quality, capacity and cost in the perioperative process at massachusetts general hospital. Master's thesis, Massachusetts Institute of Technology, 2011.

[109] J Ross Quinlan. Bagging, boosting, and c4. 5. In *AAAI/IAAI, Vol. 1*, pages 725–730, 1996.

[110] Ashleigh Royalty Range. Improving surgical patient flow through simulation of scheduling heuristics. Master's thesis, Massachusetts Institute of Technology, 2013.

[111] Wendi Rieb. Increasing patient throughput in the mgh cancer center infusion unit. Master's thesis, Massachusetts Institute of Technology, 2015.

[112] Kenneth J Rothman. No adjustments are needed for multiple comparisons. *Epidemiology*, 1(1):43–46, 1990.

[113] Susana Rubio, Eva Díaz, Jesús Martín, and José M Puente. Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. *Applied Psychology*, 53(1):61–86, 2004.

[114] G Rupert Jr et al. *Simultaneous statistical inference*. Springer Science & Business Media, 2012.

[115] Will Ruth and Thomas Loughin. The effect of heteroscedasticity on regression trees. *arXiv preprint arXiv:1606.05273*, 2016.

[116] Thomas Daniel Sanderson. Pooling and segmentation to improve primary care prescription management. Master's thesis, Massachusetts Institute of Technology, 2014.

[117] Henry Scheffe. *The analysis of variance*, volume 72. John Wiley & Sons, 1999.

[118] William R Schucany and William H Frawley. A rank test for two group concordance. *Psychometrika*, 38(2):249–258, 1973.

[119] Trevor A Schwartz. Improving surgical patient flow in a congested recovery area. Master's thesis, Massachusetts Institute of Technology, 2012.

[120] Sam Scott and Stan Matwin. Feature engineering for text classification. In *ICML*, volume 99, pages 379–388, 1999.

[121] Allan Stewart-Oaten. Rules and judgments in statistics: three examples. *Ecology*, 76(6):2001–2009, 1995.

[122] Michael W Temple, Christoph Ulrich Lehmann, Daniel Fabbri, et al. Natural language processing for cohort discovery in a discharge prediction model for the neonatal icu. *Applied clinical informatics*, 7(1):101–115, 2016.

[123] Sandra M Terra. An evidence-based approach to case management model selection for an acute care facility: Is there really a preferred model? *Professional case management*, 12(3):147–157, 2007.

[124] Terry M Therneau and Elizabeth J Atkinson. "An introduction to recursive partitioning using the rpart routines". `https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf`, 2015. [Online; accessed 20150929].

[125] Ivan Tomek. Two modifications of cnn. *IEEE Trans. Systems, Man and Cybernetics*, 6:769–772, 1976.

[126] Teresa M Treiger and Ellen Fink-Samnick. Collaborate©: a universal competency-based paradigm for professional case management, part i: introduction, historical validation, and competency presentation. *Professional case management*, 18(3):122–135, 2013.

[127] John W Tukey. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114, 1949.

[128] Gary M Weiss. Foundations of imbalanced learning. *H. He, & Y. Ma, Imbalanced Learning: Foundations, Algorithms, and Applications*, pages 13–41, 2013.

[129] Gary M Weiss and Foster Provost. Learning when training data are costly: the effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 19:315–354, 2003.

[130] Casey Whitelaw, Navendu Garg, and Shlomo Argamon. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 625–631. ACM, 2005.

[131] Adam Wright, Allison B McCoy, Stanislav Henkin, Abhivyakti Kale, and Dean F Sittig. Use of a support vector machine for categorizing free-text notes: assessment of accuracy across two institutions. *Journal of the American Medical Informatics Association*, 20(5):887–890, 2013.

[132] Stephen T Wu, Young J Juhn, Sunghwan Sohn, and Hongfang Liu. Patient-level temporal aggregation for text-based asthma status ascertainment. *Journal of the American Medical Informatics Association*, 21(5):876–884, 2014.

[133] Hui Yang. Automatic extraction of medication information from medical discharge summaries. *Journal of the American Medical Informatics Association*, 17(5):545–548, 2010.

[134] Hui Yang, Irena Spasic, John A Keane, and Goran Nenadic. A text mining approach to the prediction of disease status from clinical discharge summaries. *Journal of the American Medical Informatics Association*, 16(4):596–600, 2009.

[135] Emily Chuanmei You, David Dunt, and Colleen Doyle. How do case managers spend time on their functions and activities? *BMC health services research*, 16(1):1, 2016.

[136] Gloria Young, Lyubov Zavelina, and Vallire Hooper. Assessment of workload using nasa task load index in perianesthesia nursing. *Journal of PeriAnesthesia Nursing*, 23(2):102–110, 2008.