Fall 10-1-2021

Characterization of Postoperative Recovery After Cardiac Surgery- Insights into Predicting Individualized Recovery Pattern

Makoto Mori
Yale University Graduate School of Arts and Sciences, makoto.mori@gmail.com

Abstract

Characterization of Postoperative Recovery After Cardiac Surgery

Insights into Predicting Individualized Recovery Pattern

Makoto Mori

2021

Understanding the patterns of postoperative recovery after cardiac surgery is important from several perspectives: to facilitate patient-centered treatment decision making, to inform health care policy targeted to improve postoperative recovery, and to guide patient care after cardiac surgery. Our works aimed to address the following: 1) to summarize existing approaches to measuring and reporting postoperative recovery after cardiac surgery, 2) to develop a framework to efficiently measure patient-reported outcome measures to understand longitudinal recovery process, and 3) to explore ways to summarize the longitudinal recovery data in an actionable way, and 4) to evaluate whether addition of patient information generated through different phases of care would improve the ability to predict patient's outcome.

We first conducted a systematic review of the studies reporting on postoperative recovery after cardiac surgery using patient-reported outcome measures. Our systematic review demonstrated that the current approaches to measuring and reporting recovery as a treatment outcome varied widely across studies. This made synthesis of collective knowledge challenging and highlighted key gaps in knowledge, which we sought to address in our prospective cohort study.

We conducted a prospective single-center cohort study of patients after cardiac surgery to measure their recovery trajectory across multiple domains of recovery. Using a digital platform, we measured patient recovery in various domains over 30 days after surgery to visualize a granular evolution of patient recovery after cardiac surgery. We used a latent class analysis to facilitate

identification of dominant trajectory patterns that had been obscured in a conventional way of reporting such time-series data using group-level means. For the pain domain, we identified 4 trajectory classes, one of which was a group of patients with persistently high pain trajectory that only became distinguishable from less concerning group after 10 days. Therefore, we obtained a potentially actionable insights to tailoring individualized follow-up timing after surgery to improve the pain control.

The prospective study embodied several important features to successfully conducting such studies of patient-reported outcomes. This included the use of digital platform to facilitate efficient data collection extending after hospital discharge, iteratively improving the protocol to optimize patient engagement including evaluation of potential barriers to survey completion, and using latent class analysis to identify dominant patterns of recovery trajectories. We outlined these insights in the protocol manuscript to inform subsequent studies aiming to leverage such a digital platform to measure longitudinal patient-centered outcome.

Finally, we evaluated the potential value of incorporating health care data generated in the different phases of patient care in improving the prediction of postoperative outcomes after cardiac surgery. The current standard of risk prediction in cardiac surgery is the Society of Thoracic Surgeons' (STS) risk model, which only uses patient information available preoperatively. We demonstrated through prediction models fitted on the national STS risk model for coronary artery bypass graft surgery that the addition of intraoperative variables to the conventional preoperative variable set improved the performance of prediction models substantially. Using machine learning approach to such a high-dimensional dataset proved to be marginally important. This work demonstrated the potential value and importance of being able to leverage health care data to continuously update the prediction to inform patient outcomes and guide clinical care.

Our work collectively advanced knowledge in several key aspects of postoperative recovery. First, we highlighted the knowledge gap in the existing literature through characterizing the variability in

the ways such studies had been conducted. Second, we designed and described a framework to measure postoperative recovery and an analytical approach to informatively characterize longitudinal patient recovery. Third, we employed these designs in a prospective cohort study to measure and analyze recovery trajectories and described clinical insights obtained from the study. Finally, we demonstrated the potential value of a dynamic risk model to iteratively improve its predictive performance by incorporating new data generated as the patient progresses through the phase of care. Such a platform has the potential to individualize patient's post-acute care in a data-driven manner.

Characterization of Postoperative Recovery After Cardiac Surgery

Insights into Predicting Individualized Recovery Pattern

A Dissertation
Presented to the Faculty of the Graduate School
Of
Yale University
In Candidacy for the Degree of
Doctor of Philosophy

By
Makoto Mori

Dissertation Director: Harlan M. Krumholz, MD SM

December, 2021

Thesis Committee

Harlan M. Krumholz, MD SM

Harold H. Hines, Jr. Professor of Medicine

Department of Internal Medicine, Section of Cardiovascular Medicine

Yale School of Medicine

Director, Yale-New Haven Hospital Center for Outcomes Research and Evaluation

Arnar Geirsson, MD

Professor of Surgery (Cardiac)

Department of Surgery, Division of Cardiac Surgery

Yale School of Medicine

Sarwat Chaudhry, MD

Associate Professor of Medicine (General Medicine)

Associate Dean for Student Research, Office of Education

Yale School of Medicine

Co-Director, National Clinician Scholars Program

Yawei Zhang, MD MPH PhD

Chair, Department of Cancer Prevention and Control

China National Cancer Center

Professor, Peking Union Medical College

Acknowledgements

To Dr. Harlan Krumholz: Having the opportunity to learn from you is one of the best things that happened in my career. I am truly grateful for your mentorship and guidance. I am grateful for many insights you shared with me, not only on academic matters but also on the personal life and career development. The skillsets and the way of thinking that you provided me through working on projects have drastically changed the trajectory of my career. I cannot wait to see the opportunity resulting from the incredible training that you provided me with, and I look forward to continuing working together with you.

To Dr. Arnar Geirsson: Thank you for your unwavering support to pursue a Ph.D. during my clinical training. I am grateful for the opportunity to pursue this dedicated academic training during residency. Thank you for always supporting my academic interests and research ideas, which was critical to fostering my early interest in outcomes research.

I am looking forward to furthering my surgical training under your guidance.

To Dr. Sarwat Chaudhry: I am grateful for your mentorship to consider applying for the Investigative Medicine Program. Your generosity resulted in a key moment that pushed me to pursue this path. I am also thankful for the National Clinicians Scholars Program that you direct, from which I gained a strong background in fundamental tools needed to become a successful clinical researcher. Learning from working with your SILVER-AMI data also provided me with invaluable insights on execution and implementation and inspired me to pursue a large multicenter study at some point in my career.

To Dr. Yawei Zhang: In addition to your guidance on statistical analysis, I am grateful for your willingness to collaborate on many projects with me starting during my early research career, even before my

enrollment in the program. Your teaching and generosity to work with my ideas allowed me to pursue a career in research further.

To my collaborators: I am grateful for the expertise that my collaborators were willing to lend me and teach me along the way. Dr. Heather Allore, I am extremely thankful for your generosity in teaching me the principles, implementation, and interpretation of the latent class analysis. Your guidance elevated the quality of the analyses and inferences tremendously. Dr. Sanket Dhruva, thank you for your tireless work in forming the foundation of implementing electronic patient-reported outcome data collection. Your work and guidance in the study design were invaluable in the success of launching the prospective study.

To the leadership of the Investigative Medicine Program, Dr. Craft and Dr. Shapiro: Thank you for your effort in organizing and running this training program. I am thankful for the excellent program and rare opportunity that allows clinicians to pursue graduate degrees during residency.

I dedicate this work to my wife, ***Kayoko,*** *and our sons,* ***Yuma*** *and* ***Sota.*** *Kayoko, how you cherished and enjoyed your journey as a Ph.D. student inspired me to pursue this path. You are an amazing academic, wife, and mother.*

Table of Contents

CHAPTER 1

Characterizing Patient-Centered Postoperative Recovery After Adult Cardiac Surgery: A Systematic Review

Introduction

Postoperative recovery is a complex, time-dependent process with multiple relevant domains, including physiological, nociceptive, mental health, cognitive, sleep, mobility, and activity of daily living.[1-3] Understanding postoperative recovery after cardiac surgery is pertinent as there is increasing emphasis on readmission and outcomes of post-acute care, with implementation of national publicly-reported measures and incentive systems, such as bundled payments and Hospital Readmissions Reduction Program.[4, 5] There are increasing calls for the use of patient-reported outcomes measures (PROMs) to improve recovery, as well as digital health tools to assess function and activity.[3, 6] In fact, the Centers for Medicare and Medicaid Services is now paying for such remote monitoring.[7] However, the quality and volume of the evidence base guiding this effort in cardiac surgery population are unknown.

To inform strategies to study and improve postoperative recovery, it is important to systematically evaluate the volume, quality and content of existing literature. Of particular interest is the use of standardized methods to assess various domains relevant to recovery and inclusion of diverse patient population. Additionally, characterizing approaches to reporting PROM scores is important, as

variable reporting of raw measured scores, relative change from the preoperative measurements, or other ways, may impede generalizable synthesis of the literature. However, to date, there is no extensive review of the magnitude and quality of the studies, how prior studies have used PROM instruments and what patient populations are being studied.

Accordingly, we performed a systematic review in order to 1) describe the methods used in existing studies that evaluated postoperative recovery after cardiac surgery using PROMs, and 2) assess the populations studied. The findings will help prioritize future research by identifying areas of postoperative recovery that currently lacks data.

Methods

Search Strategy and Study Selection

We developed the protocol according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement[8]. In order to identify prospective studies on cardiac surgical population that evaluated postoperative recovery using PROMs, publications were searched on Medline and Web of Science using a combination of key terms and index headings related to cardiac surgery and postoperative recovery. We consulted a librarian experienced in systematic review on methodology and refining search terms. We did not include specific PROM terms or domain terms in order to increase the search sensitivity. We reviewed all publications indexed through January 10, 2019. The list of MeSH terms (permutations of 'postoperative,' 'cardiac surgery,' and 'recovery') and other search strategies are outlined in Supplementary Text. We reviewed search results to confirm inclusion of 5 validation articles[9-14] that we identified before the search.

We included only prospective studies in adult patients (age 18 years or older) who underwent any type of cardiac surgery that reported any PROMs following surgery. We excluded case reports and review articles. We excluded studies including patients who underwent left ventricular assist device,

extracorporeal membrane oxygenation support, orthotopic heart transplant, and congenital or adult congenital patients, as these populations likely experience distinct recovery trajectories different from the majority of adult cardiac surgical populations, which are those undergoing coronary artery bypass grafting, valve, and aortic operations. We also excluded studies with follow-up durations of fewer than 4 post-operative days, as the aim of this study was to characterize the recovery beyond acute phase of the care. To focus on studies evaluating patient-centered recovery, we excluded studies not reporting PROMs, with the exception of studies measuring physical function using accelerometers. Additionally, studies measuring PROMs at unspecified time points were not included. We added this criterion to exclude studies that obtained PROMs at undefined time points from the index operation, which can have considerable time range and is challenging to interpret considering the time-dependent nature of recovery.

Screening and data collection

We organized the articles using Endnote 8 (Clarivate Analytics, Philadelphia, PA) and two authors (MM and SA) screened the titles and abstract of all search results to locate potentially eligible articles for full-text review. Both authors then reviewed the full-text to identify the final list of eligible articles, and all disagreements were resolved by consensus.

Data Extraction

For each article, we recorded publication characteristics (first author, year of publication, and journal), study characteristics (instruments used to evaluate recovery, such as SF-36,[15] Quality of Recovery score,[16] or battery of neurocognitive tests, number of assessments performed, longest time of patient follow-up, timing of each follow-up in terms of days since the operation, the domains of recovery evaluated, inclusion/exclusion criteria, enrollment approach, missing data treatment, and how death during the follow-up was analyzed), and patient characteristics (age, sex, race, number of patients in the study, and cardiac surgery type). Patient follow-up duration was defined as the duration between

the operation and the time when the latest PROM recording was obtained. Values for the timing of measurement were collected in days since the operation. In order to assign a numeric value for visual representation of when the measurements were taken, the timing of measurement obtained at hospital discharge were defaulted to postoperative day 7, if the study did not report specific timing of postoperative discharge. Day 7 was chosen based on the mean postoperative length-of-stay of 6.9 days reported by the national Society of Thoracic Surgeons database for patients undergoing isolated CABG[17]. Journal type was grouped into 6 categories: nursing, surgical, psychology/behavioral, anesthesia, cardiology, and other. We categorized journals based on the journal title including the name of the specialty (e.g., anesthesiology, nursing) and professional society's affiliations to the journal (Supplementary table S1).

PROM domains

Six domains that characterize postoperative recovery were identified based on a previous literature review[3]: Nociceptive symptoms, physical function, activity of daily living (ADL), sleep, cognitive function, and mental health domains. Depression, anxiety, and psychosocial function were categorized into the mental health domain. Nociceptive symptoms domain included reporting of pain, physical discomfort, shortness of breath, and nausea. The physical function domain included measurement obtained either using objective tools, such as accelerometer, or PROMs. This criterion was set in order to not exclude studies that used a more rigorous tool to measure the domains. Similarly, studies using polysomnography for sleep were included to capture studies on postoperative sleep pattern, although polysomnography is likely not applicable for clinical home monitoring.

Definition of outcomes reporting methodology

In order to evaluate how PROM values are analyzed and reported, we categorized reportings into the following 7 categories: raw score, percent of patients with or without symptoms or dysfunction (according to each study's definition of categorizations), difference from baseline values, percent of

patients achieving baseline values, frequency of symptoms, fitting a model over raw scores, and others. To the best of our knowledge, there is no existing categorization of PROM reporting for postoperative period. Therefore, we identified common reporting patterns by (1) reviewing the reporting of all included studies, (2) defining major categories, and (3) conducted a second review to categorize the studies by reporting approaches. Raw score indicates reporting of mean/median value of the PROM score obtained at given time point, and represents the simplest form of reporting. All other reporting categories involve processing of the raw score, such as calculating relative changes from baseline, or proportion of the patients reaching the baseline value at given time points.

Patient characteristics, enrollment approach, and inclusion/exclusion criteria

We then evaluated demographic data, enrollment approach, and inclusion/exclusion criteria to characterize the breadth of patient populations studied. Enrollment approach was categorized into convenience sampling, consecutive enrollment, or unspecified. Inclusion and exclusion criteria of interests were those specifically outlining age, sex, comorbidity criteria, and whether studies excluded patients based on case acuity status (elective vs. non-elective).

Treatment of death and missing data

Lastly, we evaluated how patients who died during the follow-up period were treated in the analysis, in order to understand common analytical practice and existing knowledge of recovery process prior to death. To characterize potential bias due to missing data, we recorded how missing data were being handled, because in longitudinal studies with decline in study participation over time, the population retained to the completion of the study may represent a biased cohort[18].

Analysis

Studies were summarized using descriptive statistics by the sample size, procedure types, duration and timing of follow-up, number of measurements obtained, and the number of domains evaluated. Each variable was summarized either by the percentage or by the median, interquartile range (IQR), and range. Distributions of the studies in each component were summarized in a bubble plot. The most frequently used PROM instruments were selected to visualize the timing in days from operation and frequency of measurements obtained.

Results

Selected studies

The search criteria yielded 3,432 studies that potentially addressed postoperative recovery after cardiac surgery. Title and abstract screening excluded 3,267 studies. Common reasons for exclusion included studies addressing congenital heart disease population, animal studies, and studies not assessing PROMs. The remaining 165 potentially eligible articles underwent full-text review. This process excluded an additional 60 studies, consisting of studies with measures obtained at inconsistent time points, studies without full text, follow-up duration < 4 days, and those evaluating the same study sample used in other included publications. Finally, 105 articles were included for analyses (Figure 1).

Study characteristics

For the 105 included articles, the sample size of the studies tended to be small with median of 119 patients (IQR 62-229, range 14-7,321). Thirty-five percent (n=37) of the studies were intervention-based, comparing recovery between specific intervention and control groups. Twenty-five percent of the studies (n=26) were randomized controlled clinical trials, in all of which the interventions were hypothesized to improve recovery, including less-invasive surgical approach[19] and the use of special undergarments for women's incisional discomfort.[20] Seventy-seven percent (n=81) were conducted in

single-center settings. Median follow-up duration was 91 (IQR 42-182) days. Frequent follow-ups (measurements at ≥5 time points) were obtained in 15% (n=15). Studies most commonly assessed 1 domain (n=42, 40%). The nociceptive symptom domain was the most commonly measured (n=60, 57%), followed by the mental health (n=58, 55%) domain. One study that met the inclusion criteria evaluated postoperative taste change,[21] which did not meet any of our pre-specified domain categories (Table 1). Of note, studies with the largest sample size (N=7,321) evaluated only one domain, with one study having only two follow-ups[22] while another having 7 follow-ups but spanning only for 7 days (Study 17, Supplementary Table S2).

Studies were most commonly published in nursing journals (n=30, 30%), followed by surgical journals (n=25, 24%) (Supplementary Table S3). The oldest study was published in 1980 and 88% and 40% of the included studies were published after 2000 and 2010, respectively (Supplementary Figure 1).

Reporting methodology

Of the 105 studies, 71 (68%) reported only the raw scores obtained from measurement tools. Fourteen (13%) defined presence of symptoms or dysfunction in a binary form and reported proportion of patients experiencing the symptoms or dysfunction at each time point. Ten (10%) studies reported measurement values in relation to the baseline values, either as the absolute or relative difference, or proportion of patients achieving the baseline value at each measured time points (Table 2). Only 60 (57%) studies obtained the first measurement prior to the operation (Figure 2).

Most of the studies with 1-2 follow-up assessments examined duration of less than 30-days. Three studies reported 5 measurements within 50-day period,[11, 13, 23] representing the highest temporal resolution (Figure 3 and Supplementary Figure S2).

Figure 4 summarizes the measurement timing and frequencies by the studies using 36-Item Short Form Health Survey (SF-36),[10, 13, 14, 24-42] which was the most commonly used tool among the studies analyzed. Among the studies using SF-36, the total number of measurements obtained ranged

from 1 to 6, with highly variable timing of measurements among the studies. Preoperative, 42 days (6 weeks), 91 days (3 months), and 182 days (6 months) after surgery were common time points to obtain the measurement.

Patient characteristics, selection criteria, missing values

Of the 100 studies that reported sex, men represented 71% (n=27,308) of the patients. Only 26% (n=27) of the studies reported race, and of those that reported race, Caucasian race comprised 88% (n=4,852). The most common procedure type evaluated was isolated or concomitant CABG only (n=60, 57%), followed by studies including both CABGs and other non-CABG procedures (n=38, 36%); studies focusing solely on valve surgery cohort comprised 6% of the studies (n=6). Studies commonly excluded patients who died during the follow-up period (46%) and 45% did not specified how people who died were analyzed (Table 3). Only one study evaluated recovery in relation to mortality as an outcome[22].

Over half of the studies did not specify whether enrollment was consecutive or on convenience basis. Studies commonly set criteria to select for elective cases (53%) and patients with less comorbidity (64%). Ten percent of the studies set criteria to select for older patient population (age ≥60 years), and 5% of the studies specified inclusion of women only (Table 3).

Study findings

The variability in methodologies used across studies precluded synthesis of the existing evidence. Therefore, we summarized interventions and clinical characteristics associated with postoperative recovery that studies identified (Supplementary Table S4), although interpretation of such claims are difficult in the context of limited quality of studies included in this analysis.

Discussion

In this systematic review, we identified that the body of literature on postoperative recovery after cardiac surgery is small (105 studies) and limited in quality, mostly single-center studies focusing

on narrow diversity of patients. Patients studied were predominantly men and of 26% of the studies reporting race, 88% were Caucasian. Measurement and reporting methods varied widely among the studies, with no standardized use of instruments. Although studies reported predictors of recovery, most lacked external validation, were low in quality, and limited in breadths of the population studied. A significant implication of our findings is in highlighting the need for high-quality research using a standardized the approach so that recovery can be measured and improved on evidence-based fashion, especially with the current focus on post-acute phase of care.

This review has marked implications to researchers and funding bodies, as it revealed how limited the evidence on postoperative recovery is when significant interest exists in readmission reduction and improving the quality of post-acute care. The Centers for Medicare and Medicaid Services is developing PROMs as part of its Quality Payment Program to relate patient experience to hospital reimbursement[43]. This signals the need for the science behind measuring patient experience to catch up to the practice, and that need is not being fulfilled by current literature. A major implication to clinicians is that interventions to optimize postoperative recovery are based on little evidence at this point, and drawing clinical guidance on this topic from the literature is challenging.

Measurement methodologies

Significant heterogeneity and methodological weaknesses were noted in the duration of follow-ups, the frequency of measurement, tools used to assess recovery, and the domains that were assessed. Even among 22 studies using the same SF-36 instrument, there was a high variation in when, in relation to the time of surgery, and how frequently the assessments were obtained. Because such variation complicates interpretation of the results across studies, a priority area in studying postoperative recovery may be to identify standard approach to measurement frequency and timings. In addition, although accounting for individual variations in preoperative level of measurement may be important to

contextualize postoperative recovery, measurement of preoperative values was inconsistent, with only 57% of the studies performing preoperative measurement. Furthermore, the review highlighted the low temporal granularity in measurement, with 8 and 9 being the highest numbers of measurements obtained over a relatively long period of 6 months[12] to 1 year[44]. Because digital platforms may allow for a high-frequency measurement of PROM, as frequent as on a daily basis[45], leveraging such technology provides novel opportunities to obtain granular insights into the process of recovery.

Reporting methodologies

Reporting of PROMs varied across studies, representing another element that requires standardization to promote cohesive interpretation of the evidence. A majority of the studies (67%) reported results as raw scores, often as the group-level mean or median and standard deviations, without any further processing of the score. Other studies sought to provide more clinically intuitive values, such as the proportion of patients reaching the preoperative values in the measured domains or items.[9, 10] Defining the recovery as the time that one reaches preoperative level of function in each of global domains[3, 9] may be useful in the clinical setting in providing estimate of the time it takes for certain proportion of the cohort to achieving 'recovery'. However, this approach to reporting may not be as useful in assessing domains that do not have a clear improving or declining trajectory, such as the mental health domain,[10] and is also not possible when the preoperative (baseline) values are not measured. Additionally, the binary categorization of the scores limits the interpretation of recovery to that at the group-level, and obscures distributional properties, such as the standard deviation, of the raw scores. Furthermore, improvement of scores beyond baseline are not reflected in this reporting

Raw scores measured via instruments calibrated to certain population-based distributions may be difficult to interpret in highly selective cohort such as those recovering after cardiac surgery, because the clinical characteristics of specific subpopulations may not match that of the population from which

the calibration was obtained. SF-36 score was linearly transformed to have the mean score of 50 and standard deviation of 10,[46] and has been validated by the original authors across 24 patient populations with variable sociodemographic characteristics and disease severity.[47] However, whether this norm holds true in a highly specific subpopulations, such as a postoperative cohort after high acuity operations recovering from a critical care setting, is uncertain. Taken together, standardization of reporting is needed, which may entail reporting of both raw scores obtained by the instruments and any post-processing of the scores if they provide additional interpretive advantages.

Underrepresented population

We identified underrepresented populations in this review. As the vast majority of the studies (92%) selected for CABG or mixture of CABG and other operations, existing data on postoperative recovery after non-CABG operations are limited. Only 6 studies exclusively evaluated valve operations. Because the mortality and complication incidences vary across case types[48], the process of recovery is expected to also vary and likely represent an important area of investigation. Expectedly, non-Caucasian and female patients were underrepresented but more importantly, only 26% of the studies reported race data. Recovery process is reported to be more protracted in female patients[14], and racial differences in recovery and the underlying causes likely warrant investigation. Most studies excluded or did not specify the treatment of mortality that occurred during the follow-up. While exclusion may be a practical approach to handling missing data, excluding deceased patient leaves the trajectory or recovery prior to death unknown. Similarly, a large number of studies excluded patients undergoing non-elective cases with higher comorbidity levels and enrolled patients on convenience basis. Although such approaches may improve response rates, they obscure the recovery process of sicker patients. Measuring recovery of this population requires patient engagement and creatively devising ways to simplify patient response.

Design and domain

The use of objective mobility tracker device in this population was infrequent (3 studies). As the prognostic value of objectively-measured mobility has been demonstrated in oncologic[49] and non-cardiac surgical populations,[6] it may be an important aspect of global recovery assessment. Sleep and cognitive domains represented the least frequently assessed domains, although both domains undergo significant disturbances postoperatively[50, 51]. This relative infrequency may owe to the challenge related to resource-intensive cognitive function testing and polysomnography being the gold standards[52]. In order to generate evidence in a large cohort representing wide spectrum of patient population, the use of subjective surrogate measures, such as self-perceived sleep quality and duration, may be a practical alternative.

Limitations

This systematic review should be interpreted in the context of several potential limitations. First, the analysis was dependent on the available published data and is limited by publication bias and applicability of historical publications to contemporary clinical and research practice. However, we evaluated the temporal trend in the publication of included studies to assess contemporariness, and found that almost 90% of the eligible studies were published after year 2000. Second, although we worked with an experienced librarian to define the inclusive search terms and searched two large databases, it is possible that relevant studies may not have been identified. Third, the heterogeneity of studies in methodology and reporting precluded meta-analysis. We reported qualitative summary of the studies in the form of predictors of recovery reported. Fourth, although a systematic review typically include risk of bias assessment, this study focused on the synthesis of meta-data of broad types of studies, and the heterogeneity of study types precluded systematic assessment of risk of bias applicable to all studies. As the main aim of the study was to describe the characteristics of all existing studies on

this topic, we believe the metrics we used to characterize the studies provide a unified view of existing literature.

Conclusions

Our systematic review on post-operative, patient-centered outcomes after adult cardiac surgery revealed that studies are quite limited in what they assess, most often single site without external validation, varied in their approach to missing data, and narrow in terms of the diversity of patients. The evidence base regarding post-operative patient-centered outcomes needs to be strengthened in order to guide data-driven improvement of post-operative recovery. Priority areas include augmenting the volume and quality of studies, improving and standardizing the methods and PROM instruments, and focused recruitment of minority populations.

References

1. Moore FD. Getting well: the biology of surgical convalescence. *Ann N Y Acad Sci*. 1958;73:387-400.
2. Lee L, Tran T, Mayo NE, Carli F and Feldman LS. What does it really mean to "recover" from an operation? *Surgery*. 2014;155:211-6.
3. Bowyer AJ and Royse CF. Postoperative recovery and outcomes--what are we measuring and for whom? *Anaesthesia*. 2016;71 Suppl 1:72-7.
4. Wadhera RK, Yeh RW and Joynt Maddox KE. The Rise and Fall of Mandatory Cardiac Bundled Payments. *JAMA*. 2018;319:335-336.
5. Khera R, Dharmarajan K, Wang Y, Lin Z, Bernheim SM, Normand ST and Krumholz HM. Association of the Hospital Readmissions Reduction Program With Mortality During and After Hospitalization for Acute Myocardial Infarction, Heart Failure, and Pneumonia. *JAMA Netw Open*. 2018;1:e182777.
6. Daskivich TJ, Houman J, Lopez M, Luu M, Fleshner P, Zaghiyan K, Cunneen S, Burch M, Walsh C, Paiement G, Kremen T, Soukiasian H, Spitzer A, Jackson T, Kim HL, Li A and Spiegel B. Association of Wearable Activity Monitors With Assessment of Daily Ambulation and Length of Stay Among Patients Undergoing Major Surgery. *JAMA Netw Open*. 2019;2:e187673.
7. Centers for Medicare and Medicaid Services. Telehealth services. https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/TelehealthSrvcsfctsht.pdf. Accessed March 3, 2019.
8. Moher D, Liberati A, Tetzlaff J, Altman DG and Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ*. 2009;339:b2535.
9. Royse CF, Saager L, Whitlock R, Ou-Young J, Royse A, Vincent J, Devereaux PJ, Kurz A, Awais A, Panjasawatwong K and Sessler DI. Impact of Methylprednisolone on Postoperative Quality of Recovery and Delirium in the Steroids in Cardiac Surgery Trial: A Randomized, Double-blind, Placebo-controlled Substudy. *Anesthesiology*. 2017;126:223-233.
10. Diab MS, Bilkhu R, Soppa G, Edsell M, Fletcher N, Heiberg J, Royse C and Jahangiri M. The influence of prolonged intensive care stay on quality of life, recovery, and clinical outcomes following cardiac surgery: A prospective cohort study. *J Thorac Cardiovasc Surg*. 2018;156:1906-1915.e3.
11. Bowyer AJ, Heiberg J, Sessler DI, Newman S, Royse AG and Royse CF. Validation of the cognitive recovery assessments with the Postoperative Quality of Recovery Scale in patients with low-baseline cognition. *Anaesthesia*. 2018;73:1382-1391.
12. Petersen J, Vettorazzi E, Winter L, Schmied W, Kindermann I and Schäfers HJ. Physical and mental recovery after conventional aortic valve surgery. *J Thorac Cardiovasc Surg*. 2016;152:1549-1556.e2.
13. Myles PS, Hunt JO, Fletcher H, Solly R, Woodward D and Kelly S. Relation between quality of recovery in hospital and quality of life at 3 months after cardiac surgery. *Anesthesiology*. 2001;95:862-7.
14. Vaccarino V, Lin ZQ, Kasl SV, Mattera JA, Roumanis SA, Abramson JL and Krumholz HM. Gender differences in recovery after coronary artery bypass surgery. *J Am Coll Cardiol*. 2003;41:307-14.
15. Ware JE and Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care*. 1992;30:473-83.
16. Myles PS, Weitkamp B, Jones K, Melick J and Hensen S. Validity and reliability of a postoperative quality of recovery score: the QoR-40. *Br J Anaesth*. 2000;84:11-5.
17. Badhwar V, Rankin JS, Thourani VH, D'Agostino RS, Habib RH, Shahian DM and Jacobs JP. The Society of Thoracic Surgeons Adult Cardiac Surgery Database: 2018 Update on Research: Outcomes Analysis, Quality Improvement, and Patient Safety. *Ann Thorac Surg*. 2018;106:8-13.

18. Heilbrun LK, Nomura A and Stemmermann GN. The effects of non-response in a prospective study of cancer: 15-year follow-up. *Int J Epidemiol*. 1991;20:328-38.
19. Grossi EA, Zakow PK, Ribakove G, Kallenbach K, Ursomanno P, Gradek CE, Baumann FG, Colvin SB and Galloway AC. Comparison of post-operative pain, stress response, and quality of life in port access vs. standard sternotomy coronary bypass patients. *Eur J Cardiothorac Surg*. 1999;16 Suppl 2:S39-42.
20. King KM, Tsuyuki R, Faris P, Currie G, Maitland A and Collins-Nakai RL. A randomized controlled trial of women's early use of a novel undergarment following sternotomy: the Women's Recovery from Sternotomy Trial (WREST). *Am Heart J*. 2006;152:1187-93.
21. Keith M, Mokbel R, San Emeterio M, Song J and Errett L. Evaluation of taste sensitivity in patients undergoing coronary artery bypass graft surgery. *J Am Diet Assoc*. 2010;110:1072-7.
22. Koch CG, Li L, Lauer M, Sabik J, Starr NJ and Blackstone EH. Effect of functional health-related quality of life on long-term survival after cardiac surgery. *Circulation*. 2007;115:692-9.
23. Cook DJ, Thompson JE, Prinsen SK, Dearani JA and Deschamps C. Functional recovery in the elderly after major surgery: assessment of mobility recovery using wireless technology. *Ann Thorac Surg*. 2013;96:1057-61.
24. Hirschhorn AD, Richards DA, Mungovan SF, Morris NR and Adams L. Does the mode of exercise influence recovery of functional capacity in the early postoperative period after coronary artery bypass graft surgery? A randomized controlled trial. *Interact Cardiovasc Thorac Surg*. 2012;15:995-1003.
25. Morone NE, Weiner DK, Belnap BH, Karp JF, Mazumdar S, Houck PR, He F and Rollman BL. The impact of pain and depression on recovery after coronary artery bypass grafting. *Psychosom Med*. 2010;72:620-5.
26. Shuldham CM, Fleming S and Goodman H. The impact of pre-operative education on recovery following coronary artery bypass surgery. A randomized controlled clinical trial. *Eur Heart J*. 2002;23:666-74.
27. Zimmerman L, Barnason S, Hertzog M, Young L, Nieveen J, Schulz P and Tu C. Gender differences in recovery outcomes after an early recovery symptom management intervention. *Heart Lung*. 2011;40:429-39.
28. Arthur HM, Daniels C, McKelvie R, Hirsh J and Rush B. Effect of a preoperative intervention on preoperative and postoperative outcomes in low-risk patients awaiting elective coronary artery bypass graft surgery. A randomized, controlled trial. *Ann Intern Med*. 2000;133:253-62.
29. Morgan JA, Peacock JC, Kohmoto T, Garrido MJ, Schanzer BM, Kherani AR, Vigilance DW, Cheema FH, Kaplan S, Smith CR, Oz MC and Argenziano M. Robotic techniques improve quality of life in patients undergoing atrial septal defect repair. *Ann Thorac Surg*. 2004;77:1328-33.
30. Tranmer JE and Parry MJ. Enhancing postoperative recovery of cardiac surgery patients: a randomized clinical trial of an advanced practice nursing intervention. *West J Nurs Res*. 2004;26:515-32.
31. Ballan A and Lee G. A comparative study of patient perceived quality of life pre and post coronary artery bypass graft surgery. *Aust J Adv Nurs*. 2007;24:24-8.
32. Barnason S, Zimmerman L, Nieveen J, Schulz P, Miller C, Hertzog M and Rasmussen D. Relationships between fatigue and early postoperative recovery outcomes over time in elderly patients undergoing coronary artery bypass graft surgery. *Heart Lung*. 2008;37:245-56.
33. Chunta KS. Expectations, anxiety, depression, and physical health status as predictors of recovery in open-heart surgery patients. *J Cardiovasc Nurs*. 2009;24:454-64.
34. Rosenfeldt F, Braun L, Spitzer O, Bradley S, Shepherd J, Bailey M, van der Merwe J, Leong JY and Esmore D. Physical conditioning and mental stress reduction--a randomised trial in patients undergoing cardiac surgery. *BMC Complement Altern Med*. 2011;11:20.

35. Yang M, Yao M, Wang G, Xiao C, Wu Y, Zhang H and Gao C. Comparison of postoperative quality of life for patients who undergo atrial myxoma excision with robotically assisted versus conventional surgery. *J Thorac Cardiovasc Surg*. 2015;150:152-7.
36. Barnason S, Zimmerman L, Nieveen J and Hertzog M. Impact of a telehealth intervention to augment home health care on functional and recovery outcomes of elderly patients undergoing coronary artery bypass grafting. *Heart Lung*. 2006;35:225-33.
37. Körtke H, Stromeyer H, Zittermann A, Buhr N, Zimmermann E, Wienecke E and Körfer R. New East-Westfalian Postoperative Therapy Concept: a telemedicine guide for the study of ambulatory rehabilitation of patients after cardiac surgery. *Telemed J E Health*. 2006;12:475-83.
38. Miller C, Zimmerman L, Barnason S and Nieveen J. Impact of an early recovery management intervention on functioning in postoperative coronary artery bypass patients with diabetes. *Heart Lung*. 2007;36:418-30.
39. Abbott AA, Barnason S and Zimmerman L. Symptom burden clusters and their impact on psychosocial functioning following coronary artery bypass surgery. *J Cardiovasc Nurs*. 2010;25:301-10.
40. Caputo M, Alwair H, Rogers CA, Pike K, Cohen A, Monk C, Tomkins S, Ryder I, Moscariello C, Lucchetti V and Angelini GD. Thoracic epidural anesthesia improves early outcomes in patients undergoing off-pump coronary artery bypass surgery: a prospective, randomized, controlled trial. *Anesthesiology*. 2011;114:380-90.
41. Chen YC, Chen KC, Lu LH, Wu YL, Lai TJ and Wang CH. Validating the 6-minute walk test as an indicator of recovery in patients undergoing cardiac surgery: A prospective cohort study. *Medicine (Baltimore)*. 2018;97:e12925.
42. Zimmerman L, Barnason S, Brey BA, Catlin SS and Nieveen J. Comparison of recovery patterns for patients undergoing coronary artery bypass grafting and minimally invasive direct coronary artery bypass in the early discharge period. *Prog Cardiovasc Nurs*. 2002;17:132-41.
43. Centers for Medicare and Medicaid Services. CMS Quality Measure Development Plan: Supporting the Transition to theMerit-based Incentive Payment System (MIPS) and Alternative Payment Models(APMs). https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/Final-MDP.pdf. Accessed March 10, 2019.
44. Borkon AM, Muehlebach GF, House J, Marso SP and Spertus JA. A comparison of the recovery of health status after percutaneous coronary intervention and coronary artery bypass. *Ann Thorac Surg*. 2002;74:1526-30; discussion 1530.
45. Jaensson M, Dahlberg K, Eriksson M and Nilsson U. Evaluation of postoperative recovery in day surgery patients using a mobile phone application: a multicentre randomized trial. *Br J Anaesth*. 2017;119:1030-1038.
46. Gandek B, Sinclair SJ, Kosinski M and Ware JE. Psychometric evaluation of the SF-36 health survey in Medicare managed care. *Health Care Financ Rev*. 2004;25:5-25.
47. McHorney CA, Ware JE, Lu JF and Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Med Care*. 1994;32:40-66.
48. O'Brien SM, Feng L, He X, Xian Y, Jacobs JP, Badhwar V, Kurlansky PA, Furnary AP, Cleveland JC, Lobdell KW, Vassileva C, Wyler von Ballmoos MC, Thourani VH, Rankin JS, Edgerton JR, D'Agostino RS, Desai ND, Edwards FH and Shahian DM. The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 2-Statistical Methods and Results. *Ann Thorac Surg*. 2018;105:1419-1428.
49. Gresham G, Hendifar AE, Spiegel B, Neeman E, Tuli R, Rimel BJ, Figlin RA, Meinert CL, Piantadosi S and Shinde AM. Wearable activity monitors to assess performance status and predict clinical outcomes in advanced cancer patients. *npj Digital Medicine*. 2018;1:27.
50. Edéll-Gustafsson UM, Hetta JE and Arén CB. Sleep and quality of life assessment in patients undergoing coronary artery bypass grafting. *J Adv Nurs*. 1999;29:1213-20.

51. Newman MF, Kirchner JL, Phillips-Bute B, Gaver V, Grocott H, Jones RH, Mark DB, Reves JG, Blumenthal JA and Investigators NORGatCARE. Longitudinal assessment of neurocognitive function after coronary-artery bypass surgery. *N Engl J Med*. 2001;344:395-402.
52. Van de Water AT, Holmes A and Hurley DA. Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography--a systematic review. *J Sleep Res*. 2011;20:183-200.

Tables and Figures

Table 1: Study characteristics of 105 studies

Variables	N or median	% or Q1-Q3 (range)
Sample size (n)	119	62-229 (14-7321)
Randomized trial	26	25%
Intervention-based*	37	35%
Multicenter study	24	23%
Follow-up duration (days)	91	42-182 (4-1825)
Number of follow-ups		
1	7	7%
2	27	26%
3	35	33%
4	21	20%
5	7	7%
6~9	8	8%
Domains		
Nociceptive symptoms	60	57%
ADL	51	49%
Cognitive	18	17%
Mental health	58	55%
Physical function	55	52%
Sleep	11	10%
Number of domains assessed		
1	42	40%
2	14	13%
3	17	16%
4	23	22%
5	8	8%
6	0	0%

IQR= interquartile range.

*Intervention-based refers to studies that examined patient-reported outcome measures according to different process of care (robotic vs. sternotomy approach, telehealth follow-up vs. usual care, etc).

Table 2: Outcomes reporting methodology

Reporting methods	**N**	**%**
Raw score values	71	68%
Percent of patients with and without symptoms/dysfunction	14	13%
Difference from baseline	6	6%
Percent of patients achieving baseline	4	4%
Function-based (fit over raw score values)	4	4%
Frequency of symptom	3	3%
Other	3	3%

*Raw score values include 1 study reporting number of steps measured by a tracker.

Table 3: Study population characteristics

Criteria	N	%
Sex reported	100	95%
Male (of sex reported)	27,308/38,567	71%
Race reported	27	26%
Caucasian (of race reported)	4,852/5,509	88%
Procedure type		
CABG only	60	57%
CABG + other	38	36%
Valve only	6	6%
Other	1	1%
Death treatment		
Unspecified	47	45%
Excluded	48	46%
No death occurred	7	7%
Other	3	3%
Enrollment approach		
Unspecified	55	52%
Convenience	19	18%
Consecutive	31	30%
Inclusion/exclusion criteria to select for:		
Elective case only	56	53%
Non-elective case only	0	0%
Less comorbidity	67	64%
More comorbidity	4	4%
Older age (>60 years old)	10	10%
Younger age (<80 years old)	10	10%
Female sex only	5	5%

CABG= coronary artery bypass graft surgery.
Total N is 105, except for male and Caucasian numbers, which are specified in the table. Older and younger ages were defined by different thresholds in order to identify studies that focused on extremes of patient age (i.e. 'older' referred to the exclusion of extremely young population and vice versa).

Figure 1: Study selection flow chart

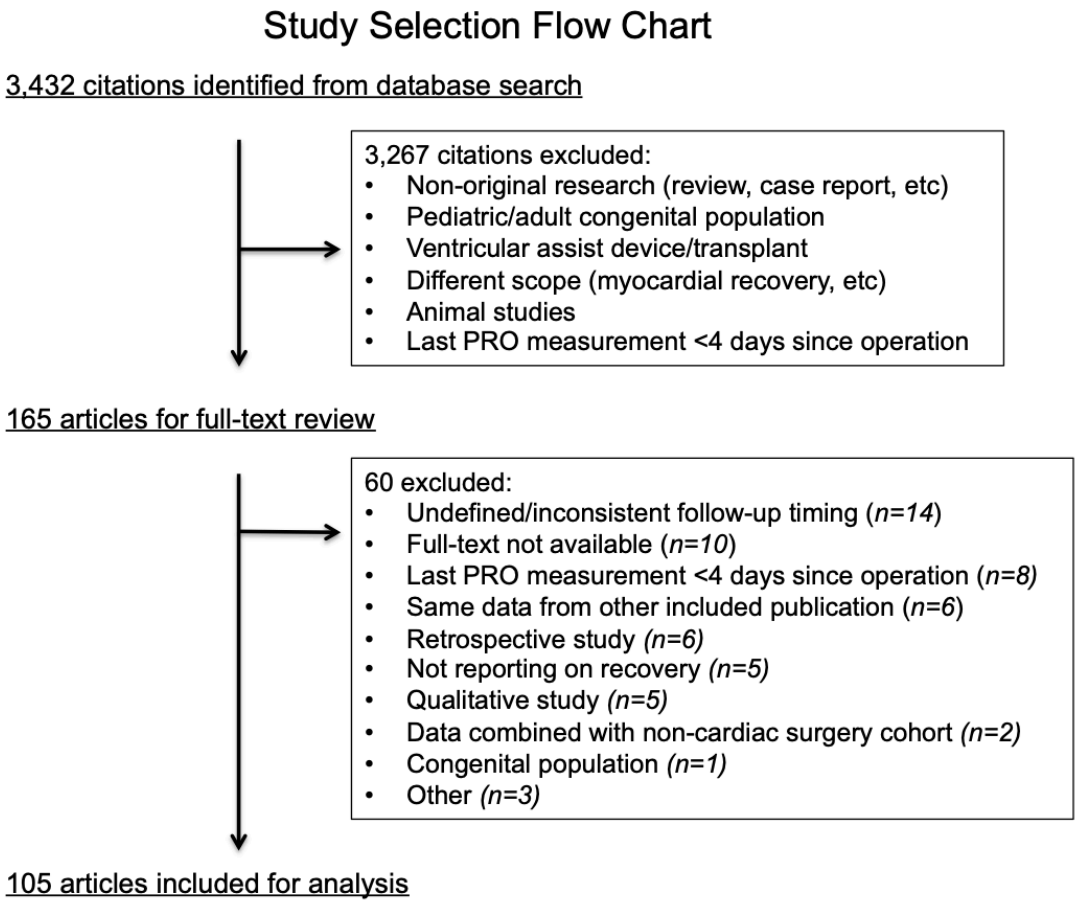
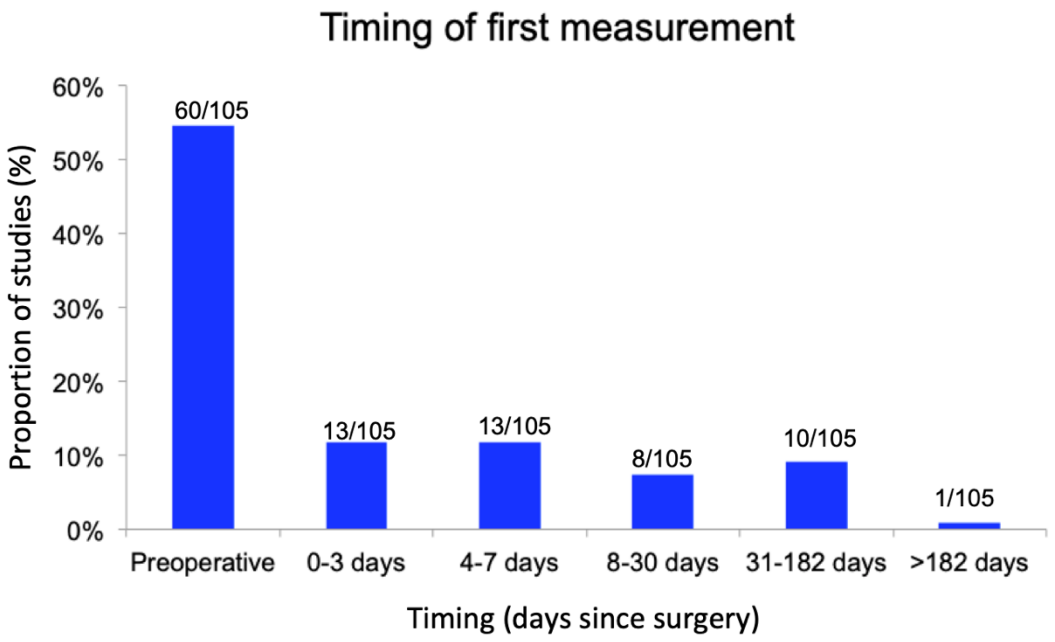


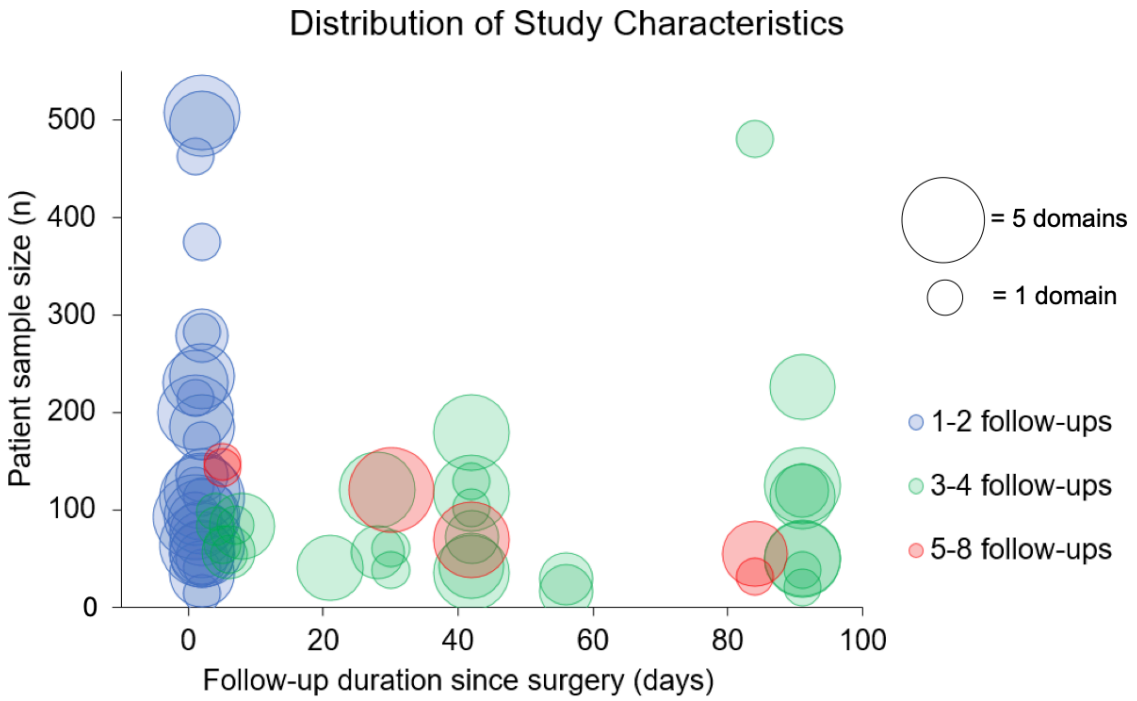
Figure shows the study selection process to arrive at the one hundred ten articles analyzed. Studies were excluded based on case types (ventricular assist device or heart transplant) and patient population (congenital, adult congenital), because the course of recovery may differ in these populations compared to common adult cardiac surgical population.
PRO= patient-reported outcomes.

Figure 2: Timing of the first measurement obtained



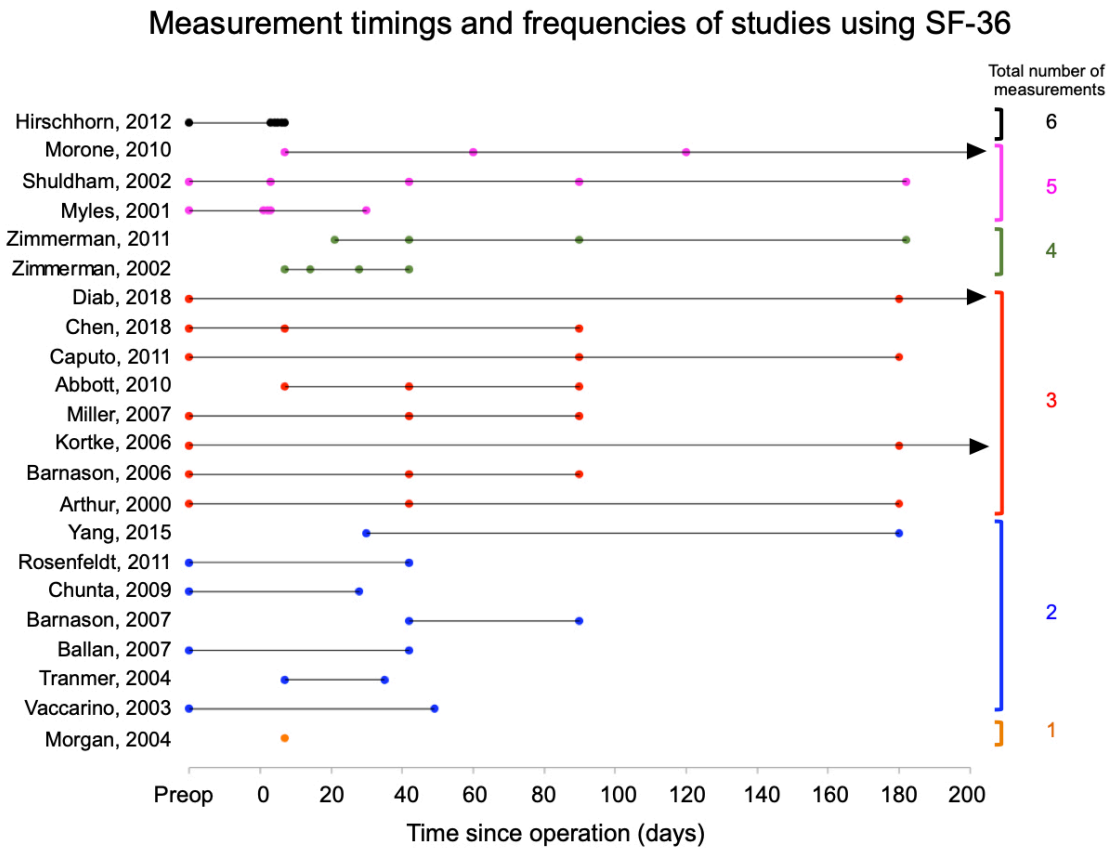
The figure displays the distribution of the timing of first measurement reported by the studies. Fifty-seven percent of the studies obtained the first measurement prior to surgery.

Figure 3: Bubble chart of studies by the study characteristics



The figure shows studies by the duration of follow-up (x-axis) up to 100 days, sample size (y-axis) up to 500 patients, number of domains evaluated (bubble size), and number of follow-ups at which time the measurements were obtained (color). Six possible domains are: nociceptive symptoms, activity of daily living, cognitive, sleep, mental health, and physical function.

Figure 4: Measurement timings and frequencies of studies using the 36-Item Short Form Health Survey (SF-36)



The figure shows measurement timing and frequencies in studies using 36-Item Short Form Health Survey (SF-36). Each horizontal line represents a study and each dot represents the time point at which measurements were obtained. Last name of the first author and publication years are displayed in the left column. Studies are clustered by the total number of measurements obtained during the study (right column). Arrows indicate follow-up > 200 days

Supplementary materials

Search terms

Medline (n=2,851):
(("postoperative period"[MeSH Terms] OR (("postoperative"[All Fields] OR "post-operative"[All Fields]) AND "period"[All Fields]) OR "postoperative period"[All Fields] OR "postoperative"[All Fields] OR “postsurgical”[All Fields]) AND recovery[All Fields] AND ("cardiac surgery"[All Fields] OR "cardiac surgical procedures"[MeSH Terms] OR "cardiac surgical"[All Fields] OR “CABG”[All Fields] OR “coronary artery bypass”[All Fields] OR “valve replacement”[All Fields] OR “valve repair”[All Fields])) AND English[Language]

Web of Science (n=1,921):
(ALL= ((Postoperative OR post-operative OR “post operative” OR postsurgical OR post-surgical OR “post surgical”) AND recovery AND ("cardiac surgery" OR "cardiac surgical" OR CABG OR 'coronary artery bypass' OR 'valve surgery' OR 'valve repair' OR valve replacement)))*AND* LANGUAGE**:** (English)

Final list after de-duplication (n=3,432)

Table S1: Journal categorization by specialty

Journals were categorized according to the inclusion of the specialty name in the journal title and the professional society that publishes the journal.

Table S2: List of included studies

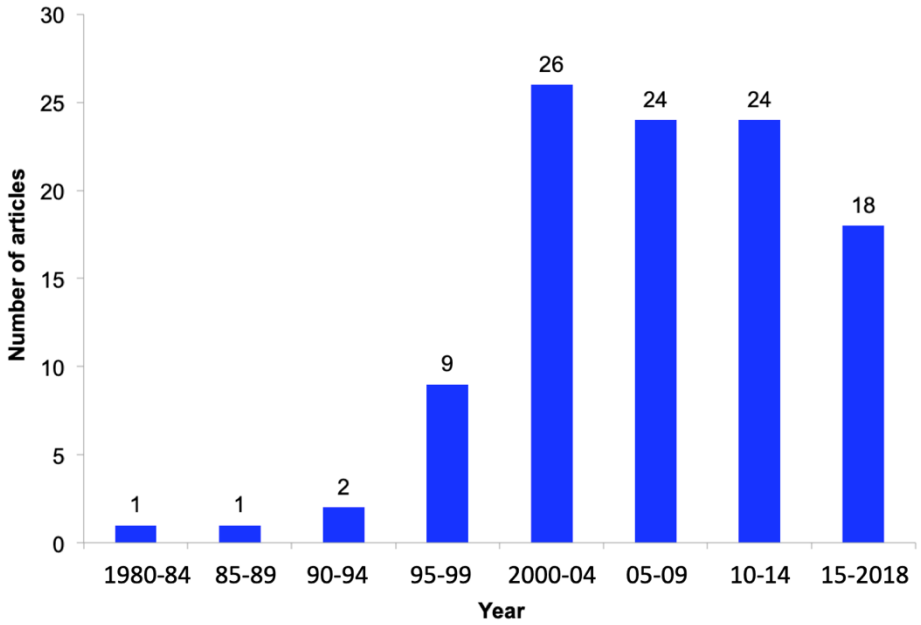
Given few studies with extremely large samples compared to the median sample size, we qualitatively described studies with the largest sample sizes to evaluate whether we could make a strong inference on normative recovery pattern. The largest study (n= 7,321), Study 45, obtained measurements at only two time points, first of which was 180 days after the operation. The second largest study (n=5,658), Study 17, obtained measurements in a single-center setting at 7 time points but only within the first 7 days, and only assessed the proportion of patients ambulating on each day.

Table S3: Number of articles by journal category

*Psych/behavioral category includes psychology, psychiatry, and behavioral medicine.

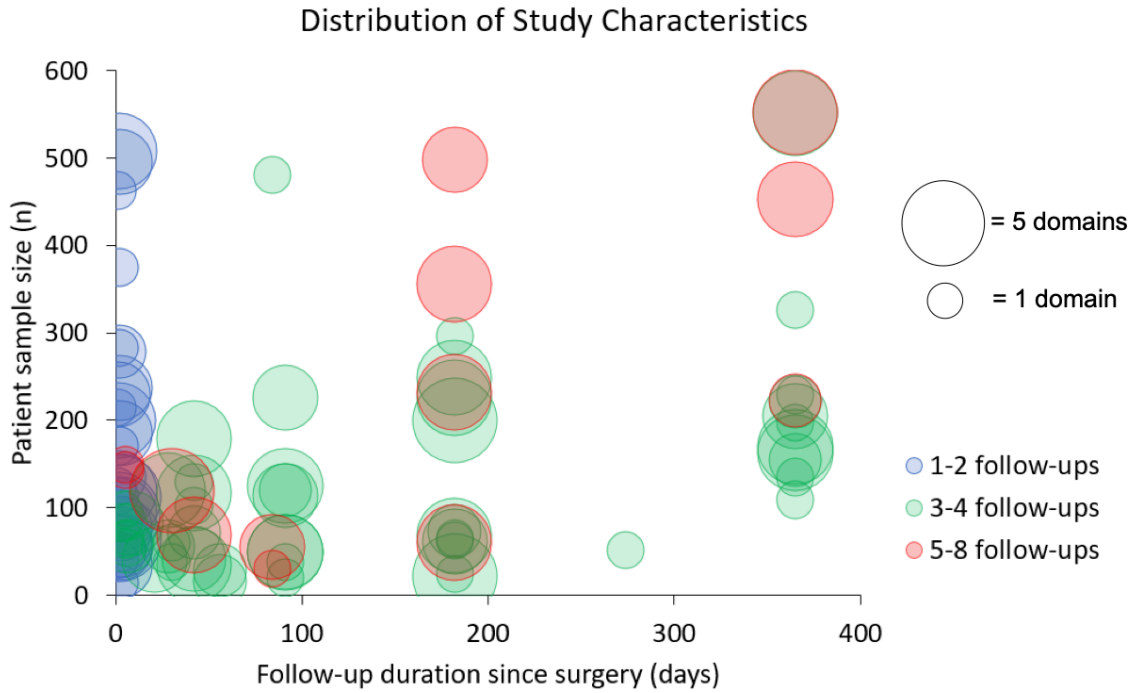
Table S4: Qualitative summary of predictors and interventions associated with improved recovery

Figure S1: Number of publications by year



The figure shows the number of articles published by 5-year increment of calendar year. No publication published prior to 1980 met the inclusion criteria. The latest bin (2015-2018) includes only 4-year period.

Figure S2: Bubble chart of studies by the study characteristics up to 400 follow-up days



The figure shows studies by the duration of follow-up (x-axis) up to 400 days, sample size (y-axis) up to 600 patients, number of domains evaluated (bubble size), and number of follow-ups at which time the measurements were obtained (color). Six possible domains are: nociceptive symptoms, activity of daily living, cognitive, sleep, mental health, and physical function.

CHAPTER 2

Characterization of Postoperative Recovery Trajectories After Cardiac Surgery

Introduction

Postoperative pain is an outcome important to patients[1] and physicians[2] as poorly controlled pain may result in prolonged rehabilitation, reduced activity level, and complications[3, 4]. Although it is expected that pain level evolves after surgery, there is great variability in how pain has been studied in terms of assessment frequency and duration[5], making it difficult to draw a collective inference. Additionally, such studies commonly report pain level as an average of the entire study cohort,[5, 6] potentially obscuring individual variations in the longitudinal experience of pain and identification of patient characteristics denoting those with either desirable or concerning progression of pain level. A more granular understanding of pain trajectories can guide judicious pain management strategies at hospital discharge and inform individualized timing of postoperative visits. Electronic platform to measure patients' pain level may enable frequent measurement even beyond hospital discharge to provide novel insight.[1]

We aimed to characterize heterogeneity of pain trajectories and explore clinical characteristics of patients with persistently low and high pain over time, using longitudinal data of patient-reported postoperative pain after cardiac surgery captured using electronic platform. We also aimed to assess how measuring responses more frequently than prior studies may relate to precisely capturing patients' pain experience.

Methods

Patient selection criteria and data source

We studied a convenience sample of patients who underwent cardiac surgery at Yale New Haven Hospital between January 2019 and March 2020. Postoperative, as opposed to preoperative, enrollment allowed us to enroll non-elective cases. Inclusion criteria were patients undergoing isolated or concomitant coronary artery bypass graft (CABG), aortic valve replacement, mitral valve replacement, mitral valve repair, or aortic operation who were discharged from the intensive care unit (ICU) within 5 days of the operation. This 5-day threshold ensured that time of initiation of pain assessments would be standardized, since patients could not be enrolled while in the ICU. Patients were enrolled with written informed consent upon discharge from the ICU after surgery. We excluded those who do not own a smartphone or a tablet or those who do not speak or read English because the electronic platform for patient-reported outcome measure (PROM) data collection relied on patients responding to surveys displayed on a web browser via email or text. Despite the need for these exclusion criteria, we chose to use the electronic platform for the automated electronic delivery of surveys that allowed for seamless collection of PROM data even after the patient's hospital discharge. We also excluded those who could not complete the enrollment process (Figure 1). Details of the protocol have been published.[7] The cardiac surgery service did not have a formalized Enhanced Recovery After Surgery pathway at the time of the study. Pain regimens were individualized to the patients' needs during the hospitalization and at the time of discharge. The Yale Institutional Review Board approved the study.

Questionnaire and data collected

Quality of Recovery (QoR-24), a 24-item questionnaire assessing postoperative recovery[8-10] adapted from the original QoR-40,[11] was emailed every 3 days for 30 days. The questionnaire item for pain read 'During the last 24 hours, I have been having pain in the

surgical wound,' with possible responses ranging from 0 to 10 with 0 corresponding to 'none of the time' and 10 corresponding to 'all of the time.' Variables describing patient characteristics were prespecified in the protocol article[7] and were collected via the institutional Society of Thoracic Surgeon's (STS) Adult Cardiac Surgery Database using the data version 2.91 definitions. Prescription of opioid medication at the time of hospital discharge was collected via chart review and standardized to morphine milligram equivalents (MME).

Statistical analysis

We evaluated the variability in pain trajectories over 30 days by visualizing the plot by 1) individual, 2) cohort-level mean, and 3) latent class group-based trajectory model. For this step, we excluded patients with fewer than 3 responses to estimate potential quadratic effects (i.e. increase and decrease). We applied a group-based trajectory model, a family of latent class analysis, which estimated the probability of belonging to a specific trajectory of pain.[12, 13] This is a semiparametric finite mixture model for longitudinal data using a maximum likelihood method fitting the pain score with a censored normal distribution. We fitted the model from one to five trajectories with polynomial order of up to a cubic term. Attrition from the study was not modeled together, as there was no mortality during the study period.

We determined the optimal number of trajectory classes based on the Bayesian information criterion and average posterior probability of assignment (>0.9 indicated excellent fit and <0.7 indicated poor fit) among the models with one to five trajectory classes and incrementally increasing the polynomial order.[14]

Representativeness and number of measurements

We tested whether increasing the number of measurements improves the representativeness of pain level. Specifically, we compared the average pain level for each

patient based on k number of measurements, against the reference of individual's average pain level over 10 measurements. This analysis was performed on 56 patients who responded to ≥7 surveys. In this subgroup, pain level was missing in 16% (93/560) of the surveys. The difference between patient-level average of pain score over 10 measurements was compared with the patient-level average of pain score over k measurements and the difference was plotted along the k measurement to visualize the relationship between the number of measurement and representativeness of the measurement.

Missing data

For the analysis of representativeness of pain measurement, we used data from those who responded to ≥7 surveys. We used a higher threshold for survey response in this analysis to minimize bias introduced by imputing missing values. We used linear interpolation to estimate the missing value for missing responses, because in this dataset, missing data were sparse across the time-series.[15] Such imputation was not used for the trajectory model that used observations with ≥3 responses, as group-based trajectory model's full information maximum likelihood estimation allowed for integration of all available information based on missing-at-random assumption.[16] Missing data for the STS data occurred in <2% of participants and missing values were conditionally estimated as described by Shahian, et al. in the STS risk model development,[17] classifying missing values to those in the lowest risk category for categorical variables and using age and sex-specific means for continuous variables.

We did not compare groups in terms of statistically significant differences because of the limited sample size. We used Traj package for a group-based trajectory model and calculated k means via Proc Univariate procedure in SAS 9.4 (SAS Institute, Inc Cary, NC). We used Python 3.8 for data preprocessing including linear interpolation.

Results

Of 92 patients enrolled, there were 75 (82%) with ≥3 responses and 56 (61%) with ≥7 responses. Characteristics of the 75 patients are summarized in Table 1. The median age of the patients was 64 (interquartile range: 58 to 70) years with 57 (76%) men, and 66 (88%) White. Thirty-four patients underwent isolated CABG, which was the most common case type.

In 75 patients with ≥3 responses, we observed that individual pain scores varied substantially across patients with no dominant mean or pattern (Figure 2, Panel A). Cohort-level mean and 95% confidence interval (Figure 2, Panel B) showed a gradual and consistent decline in the mean pain level over time, but the confidence bands covered most of the pain score range. Based on the best BIC value (Appendix Table 1), the group-based trajectory model identified 4 trajectories (Figure 2, Panel C), all of which had a posterior probability of assignment of 0.85 or higher. We labeled the trajectories according to the observed pattern: persistently low (n=9, 12%), moderate declining (n=26, 35%), high declining (n=33, 44%), and persistently high pain (n=7, 9%). Persistently high pain and high declining groups did not appear to be clearly distinguishable until the 3rd measurement of approximately 10 days.

Comparing patient, operative, and postoperative characteristics in the 4 assigned trajectories, patients in the persistently low pain trajectory class were older (numerically higher median age) than the other 3 classes. The proportions of patients who underwent robotic-assisted surgery were 1 (11%) in persistently low, 5 (19%) in moderate declining, 7 (21%) in high declining, and 3 (43%) in persistently high pain trajectory classes. Patients in the persistently high pain trajectory class had a numerically higher median length of hospital stay than the other 3 classes. The proportion of patients who were not prescribed any opioid medications at the

time of hospital discharge were 2 (22%) in persistently low, 6 (23%) in moderate declining, 6 (18%) in high declining, and 2 (29%) in persistently high pain trajectory classes.

Median morphine milligram equivalent of narcotics prescribed at the time of discharge were 90 (interquartile range [IQR] 60-100mg) in persistently low, 150 (IRQ: 60-180mg) in moderate declining, 150 (IQR: 75-150mg) in high declining, and 90 (IQR: 0-165mg) in persistently high pain trajectory classes (Table 2).

Compared with the patient-level mean of pain score over 10 measurements as a reference, the pain level determined by the first measurement alone differed by 2.1 points on average within the same patient. The pain level determined by the last measurement alone differed by 1.9 points. Increasing the number of measurements decreased the difference incrementally. Obtaining measurements every 6, 9, and 15 days for a total of 5, 3, and 2 measurements was associated with 0.5, 0.7, and 1.0 point differences from the reference, respectively (Figure 3).

Discussion

Using a longitudinal patient-reported pain measure collected up to 30 days after cardiac surgery, we identified distinct trajectories of reported pain. These result reveals that people experience very different recoveries after cardiac surgery. Revealing this heterogeneity provides an impetus to understand the determinants of different outcomes and targets of intervention to ensure that more people have less pain. It also provides the possibility of better informing patients about what the recovery experience might entail. To date, little attention has focused on quantifying the variations in the way that pain tracks through the early recovery period.

The heterogeneity of the identified pain trajectories is important because pain and other postoperative recovery domains have mostly been reported as cohort-level average even in studies that measured pain at multiple postoperative time points.[18-20] Therefore, the dominant patterns of trajectories and a heterogeneous patient experience after cardiac surgery may be obscured by focusing on average effects.[5] This study highlights the opportunity to identify factors that may be related to distinct pain trajectories. These may include factors at the patient, surgeon, and site levels. It also highlights the potential for studying heterogeneity in trajectories of other post-operative outcomes, including mobility, energy level and return to baseline activities.

This study adds to the literature on trajectories of pain following surgery in several notable ways. First, prior studies have included thoracotomy, orthopedic, and general surgery,[21, 22 23] Therefore, this study uniquely highlights pain trajectories after cardiac surgery that may differ from non-cardiac surgery due to sternotomy and less-invasive approaches. Second, prior studies focused on a much longer timescale with the follow-up at 1 year after the operation. [21, 22 23] Our granular measurement within the shorter, 30-day postoperative window may offer different opportunities for timely interventions. Given the current opioid epidemic and increased attention to judicious postoperative narcotics prescription,[24] recognizing the pain trajectory variations in the immediate postoperative period may promote more judicious and individualized narcotics prescription. Individualized opioid regimen is especially important in the postoperative period where the risk of opioid misuse is high.[2] Given the growing population of patients undergoing valve surgery due to endocarditis in the setting of injection drug use, information to guide pain management after surgery is especially relevant.[25, 26]

Our study highlights the utility of characterizing postoperative recovery and proposes an underutilized approach to measuring and reporting recovery. Because remote patient

monitoring can be reimbursed with recent expansion of the rules,[3] our study offers timely insights into the frequency of measurements needed to adequately capture patient experience. We demonstrated that relying on a single point of measurement may insufficiently represent a patient's recovery experience, and increasing the number of measurements incrementally improved the representation. In summarizing such data, the cohort-level average of pain level over time alone limits information meaningful to patients and surgeons, obscuring important variations among patients. For example, it would be less informative to patients to know that on average, pain level after cardiac surgery will mostly be between 2 to 8 points around 10 days after the operation than to know that given the first few measurements of pain, the patient is likely to have persistently high or low pain. It is important to recognize that, albeit a small study, we identified this substantial variation among this relatively homogenous patient group of younger, mostly Caucasian, and male patients.

Although the small sample size limited evaluation of associations between patient characteristics and each trajectory class of pain, it is notable that patients in the persistently high pain trajectory did not differ substantially according to opioid prescription at discharge, readmission, or sternal wound infection within 30 days. The total morphine milligram equivalent of narcotics prescribed was numerically lower compared with low or high-declining groups, but the median value of 90 was equivalent to that of persistently low pain group. Understanding associations of such potentially relevant factors and the pain trajectories require further studies.

Limitations

The single-center design of our study may limit the generalizability of our findings, although the variation in the phenotype of pain trajectories may be a finding applicable to

practices in care settings different from ours. Despite the relative homogeneity of our patient group in terms of demographics, low comorbidity profile, and low complication rates, we observed considerable variation in the trajectories of pain. Generalizability of such variation in the trajectories to different patient groups must be evaluated in future studies. The number of approached patients were lower than the number of cases usually performed at our hospital because of the COVID-19 pandemic, gradual acceleration of the enrollment rate at the beginning of the study, and the part-time availability of the research assistant to complete the enrollment. Nevertheless, the bias was toward a more homogeneous sample and yet our results reveal marked heterogeneity in patient experience even among this selected cohort. A small sample size limited our ability to make more robust inference for characteristics associated with specific recovery trajectory, including multivariable analysis and evaluation of the longitudinal change in pain regimen. As expected, many patients did not complete all 10 delivered surveys. We delivered a high number of surveys to capture at least 3 responses for the trajectory to be modeled in the latent class analysis. The reported pain levels were the perceived pain level without adjusting for variation in adherence or the prescribed narcotic dose. Therefore, this study did not evaluate whether the observed variation in pain level is associated with variation in individual pain management approaches. Nevertheless, the heterogeneity identified irrespective of the treatment approach inform potential ways to improve postoperative pain experience, including individualized timing of postoperative follow-up.

Conclusion

After cardiac surgery, the trajectory of pain is variable within 30 days. This individual variation is not adequately captured unless multiple measurements are obtained. Cohort-level mean, a common way of reporting pain level, fails to capture this variation, while a latent class

model can illustrate the heterogeneity. Studies on postoperative pain should consider the time-varying nature of pain and recognize the limitation in capturing patient experience when relying on a small number of measurements.

References:

1. Brennan F, Carr DB, Cousins M. Pain management: a fundamental human right. *Anesth Analg*. Jul 2007;105(1):205-21. doi:10.1213/01.ane.0000268145.52345.55
2. Green CR, Wheeler JR. Physician variability in the management of acute postoperative and cancer pain: a quantitative analysis of the Michigan experience. *Pain Med*. Mar 2003;4(1):8-20. doi:10.1046/j.1526-4637.2003.03006.x
3. Kehlet H, Holte K. Effect of postoperative analgesia on surgical outcome. *Br J Anaesth*. Jul 2001;87(1):62-72. doi:10.1093/bja/87.1.62
4. Sethares KA, Chin E, Costa I. Pain intensity, interference and patient pain management strategies the first 12 weeks after coronary artery bypass graft surgery. *Appl Nurs Res*. Nov 2013;26(4):174-9. doi:10.1016/j.apnr.2013.07.005
5. Mori M, Angraal S, Chaudhry SI, et al. Characterizing Patient-Centered Postoperative Recovery After Adult Cardiac Surgery: A Systematic Review. *J Am Heart Assoc*. Nov 2019;8(21):e013546. doi:10.1161/JAHA.119.013546
6. van Gulik L, Janssen LI, Ahlers SJ, et al. Risk factors for chronic thoracic pain after cardiac surgery via sternotomy. *Eur J Cardiothorac Surg*. Dec 2011;40(6):1309-13. doi:10.1016/j.ejcts.2011.03.039
7. Mori M, Brooks C, Spatz E, et al. Protocol for project recovery after cardiac surgery: a single-center cohort study leveraging digital platform to characterise longitudinal patient-reported postoperative recovery patterns. *BMJ Open*. Sep 2020;10(9):e036959. doi:10.1136/bmjopen-2020-036959
8. Jaensson M, Dahlberg K, Eriksson M, Nilsson U. Evaluation of postoperative recovery in day surgery patients using a mobile phone application: a multicentre randomized trial. *Br J Anaesth*. Nov 1 2017;119(5):1030-1038. doi:10.1093/bja/aex331
9. Halleberg Nyman M, Nilsson U, Dahlberg K, Jaensson M. Association Between Functional Health Literacy and Postoperative Recovery, Health Care Contacts, and Health-Related Quality of Life Among Patients Undergoing Day Surgery: Secondary Analysis of a Randomized Clinical Trial. *JAMA Surg*. Aug 1 2018;153(8):738-745. doi:10.1001/jamasurg.2018.0672
10. Dahlberg K, Jaensson M, Eriksson M, Nilsson U. Evaluation of the Swedish Web-Version of Quality of Recovery (SwQoR): Secondary Step in the Development of a Mobile Phone App to Measure Postoperative Recovery. *JMIR Res Protoc*. Sep 28 2016;5(3):e192. doi:10.2196/resprot.5881
11. Myles PS, Weitkamp B, Jones K, Melick J, Hensen S. Validity and reliability of a postoperative quality of recovery score: the QoR-40. *Br J Anaesth*. Jan 2000;84(1):11-5.
12. Mori M, Krumholz HM, Allore HG. Using Latent Class Analysis to Identify Hidden Clinical Phenotypes. *JAMA*. 08 2020;324(7):700-701. doi:10.1001/jama.2020.2278
13. DS N, RA T. Developmental trajectory groups: Fact or a useful statistical fiction? *Criminology*. 2005;43: 873–904.
14. Nagin D. *Group-Based Modeling of Development*. Harvard University Press; 2005.

15. Newbury J. Linear Interpolation. *Basic Numeracy Skills and Practice*. Macmillan Education UK; 1981:67-72.
16. Nagin DS, Odgers CL. Group-based trajectory modeling in clinical research. *Annu Rev Clin Psychol*. 2010;6:109-38. doi:10.1146/annurev.clinpsy.121208.131413
17. Shahian DM, O'Brien SM, Filardo G, et al. The Society of Thoracic Surgeons 2008 Cardiac Surgery Risk Models: Part 1-Coronary Artery Bypass Grafting Surgery. *Annals of Thoracic Surgery*. Jul 2009;88(1):S2-S22. doi:10.1016/j.athoracsur.2009.05.053
18. Diab MS, Bilkhu R, Soppa G, et al. The influence of prolonged intensive care stay on quality of life, recovery, and clinical outcomes following cardiac surgery: A prospective cohort study. *J Thorac Cardiovasc Surg*. Nov 2018;156(5):1906-1915.e3. doi:10.1016/j.jtcvs.2018.05.076
19. Chapman CR, Zaslansky R, Donaldson GW, Shinfeld A. Postoperative pain trajectories in cardiac surgery patients. *Pain Res Treat*. 2012;2012:608359. doi:10.1155/2012/608359
20. Milgrom LB, Brooks JA, Qi R, Bunnell K, Wuestfeld S, Beckman D. Pain levels experienced with activities after cardiac surgery. *Am J Crit Care*. Mar 2004;13(2):116-25.
21. Gjeilo KH, Oksholm T, Follestad T, Wahba A, Rustøen T. Trajectories of Pain in Patients Undergoing Lung Cancer Surgery: A Longitudinal Prospective Study. *J Pain Symptom Manage*. 04 2020;59(4):818-828.e1. doi:10.1016/j.jpainsymman.2019.11.004
22. Althaus A, Arránz Becker O, Neugebauer E. Distinguishing between pain intensity and pain resolution: using acute post-surgical pain trajectories to predict chronic post-surgical pain. *Eur J Pain*. Apr 2014;18(4):513-21. doi:10.1002/j.1532-2149.2013.00385.x
23. Bishop MO, Bayman EO, Hadlandsmyth K, Lund BC, Kang S. Opioid use trajectories after thoracic surgery among veterans in the United States. *Eur J Pain*. Jun 2020;doi:10.1002/ejp.1610
24. Brown CR, Chen Z, Khurshan F, Groeneveld PW, Desai ND. Development of Persistent Opioid Use After Cardiac Surgery. *JAMA Cardiol*. Jun 2020;doi:10.1001/jamacardio.2020.1445
25. Geirsson A, Schranz A, Jawitz O, et al. The Evolving Burden of Drug Use Associated Infective Endocarditis in the United States. *Ann Thorac Surg*. May 2020;doi:10.1016/j.athoracsur.2020.03.089
26. Mori M, Bin Mahmood SU, Schranz AJ, et al. Risk of reoperative valve surgery for endocarditis associated with drug use. *J Thorac Cardiovasc Surg*. Jul 2019;doi:10.1016/j.jtcvs.2019.06.055

Tables and Figures

Table 1: Patient characteristics

Variable	N=75 (median)	% (IQR)
Age (years)	64	(58-70)
Female	18	24%
Body mass index (kg/m^2)	28.9	(24.4-32.1)
Race		
White	66	88%
Black	4	5%
Other	5	7%
Comorbidity		
Hypertension	52	69%
Diabetes	24	32%
Stroke	8	11%
Chronic lung disease	12	16%
Peripheral vascular disease	7	9%
Liver disease	4	5%
Dialysis	2	3%
Prior cardiac surgery	5	7%
Myocardial infarction	22	29%
Heart failure	23	31%
Ejection fraction (%)	60	(55-63)
Operative details		
Non-elective cases	22	29%
Isolated CABG	34	45%
Concomitant CABG	4	5%
Aortic surgery	5	7%
Aortic valve replacement	16	21%
Mitral valve replacement	8	11%
Mitral valve repair	17	23%
Robotic approach	16	
Complications within 30 days of operation		
Mortality	0	0%
Pneumonia	1	1%
Sternal wound infection	2	3%
Pleural effusion requiring drain	4	5%
Stroke	0	0%
Renal failure	0	0%

30-day readmission	5	7%

CABG= coronary artery bypass graft.

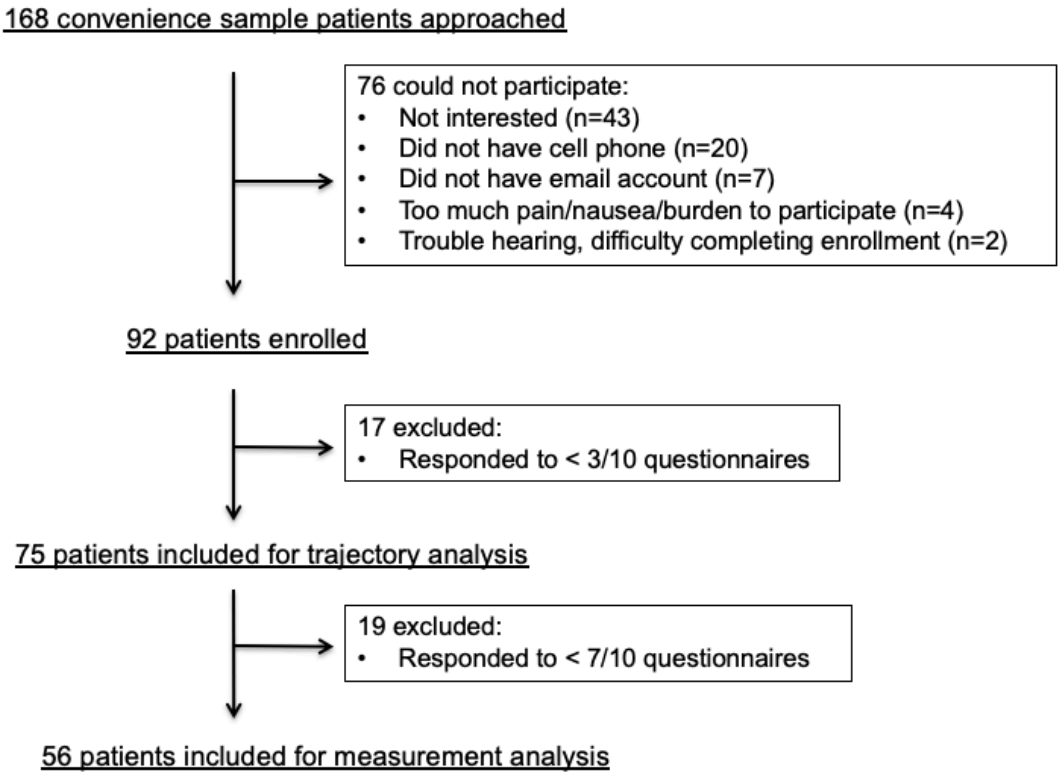
Table 2: Patient characteristics by trajectory classes

	Persistently low pain (N=9)		Moderate declining pain (N=26)		High declining pain (N=33)		Persistent
Variable	N (median)	% (IQR)	N (median)	% (IQR)	N (median)	% (IQR)	N (media
Age	68	(65-73)	65.5	(57-71)	62	(57-69)	
Female	3	33%	5	19%	7	21%	
Body mass index	29.6	(24.8-32.9)	30.5	(26.8-32.8)	27.5	(24.2-30.5)	26
Race							
White	9	100%	24	92%	27	82%	
Black	0	0%	0	0%	4	12%	
Other	0	0%	2	8%	2	6%	
Comorbidity							
Hypertension	9	100%	20	77%	19	58%	
Diabetes	4	44%	10	38%	7	21%	
Stroke	2	22%	0	0%	1	3%	
Chronic lung disease	2	22%	4	15%	6	18%	
Peripheral vascular disease	0	0%	3	12%	2	6%	
Liver disease	0	0%	2	8%	2	6%	
Dialysis	0	0%	0	0%	2	6%	
Prior cardiac surgery	1	11%	2	8%	2	6%	
Myocardial infarction	4	44%	8	31%	8	24%	
Heart failure	3	33%	7	27%	10	30%	
Ejection fraction (%)	63	(58-68)	58	(48-63)	60	(55-63)	
Pain regimen on admission							
Non-narcotic pain medication	1	11%	5	19%	7	21%	
Narcotic pain medication	0	0%	1	3%	1	3%	
Operative details							
Non-elective cases	5	56%	8	31%	6	18%	
Isolated CABG	3	33%	13	50%	14	42%	
Concomitant CABG	0	0%	1	4%	3	9%	
Aortic surgery	2	22%	2	8%	1	3%	
AVR	2	22%	5	19%	9	27%	
MVR	2	22%	1	4%	5	15%	
Mitral valve repair	2	22%	8	31%	4	12%	
Robotic assist	1	11%	5	19%	7	21%	
Non-robotic thoracotomy	1	11%	0	0%	2	6%	
Complications							

30-day mortality	0	0%	0	0%	0	0%	
Pneumonia	0	0%	0	0%	1	3%	
30-day readmission	0	0%	2	8%	3	9%	
Sternal wound infection	0	0%	1	4%	1	3%	
Pleural effusion requiring drain	0	0%	2	8%	2	6%	
Stroke	0	0%	0	0%	0	0%	
Renal failure	0	0%	0	0%	0	0%	
Postoperative length of stay	5	(4-6)	4	(4-6)	5	(4-6)	
Pain regimen at discharge							
No narcotic medication prescribed post-op	2	22%	6	23%	6	18%	
Total MME prescribed at discharge	90	(60-100)	150	(60-180)	150	(75-150)	

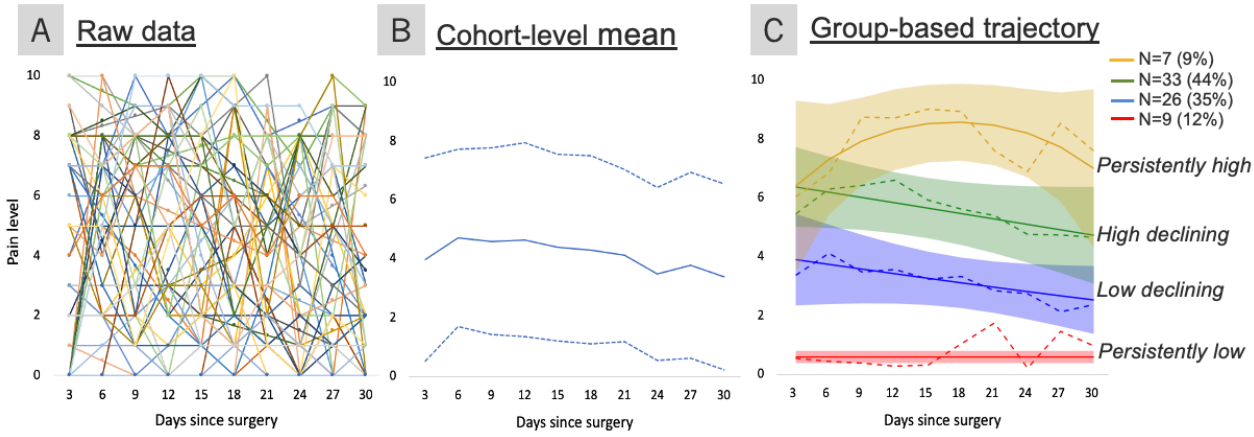
IQR = interquartile range; CABG = coronary artery bypass graft; AVR = aortic valve replacement; MVR = mitral valve replacement; MME= morphine milligram equivalent

Figure 1: CONSORT-style patient flow chart



The figure shows the flow of patient exclusion and enrollment. We included 75 patients for trajectory analysis and 56 patients for analysis evaluating the optimal interval and number of measurements. There was no mortality during the study period.

Figure 2: Various representations of the same longitudinal pain data (n=75)



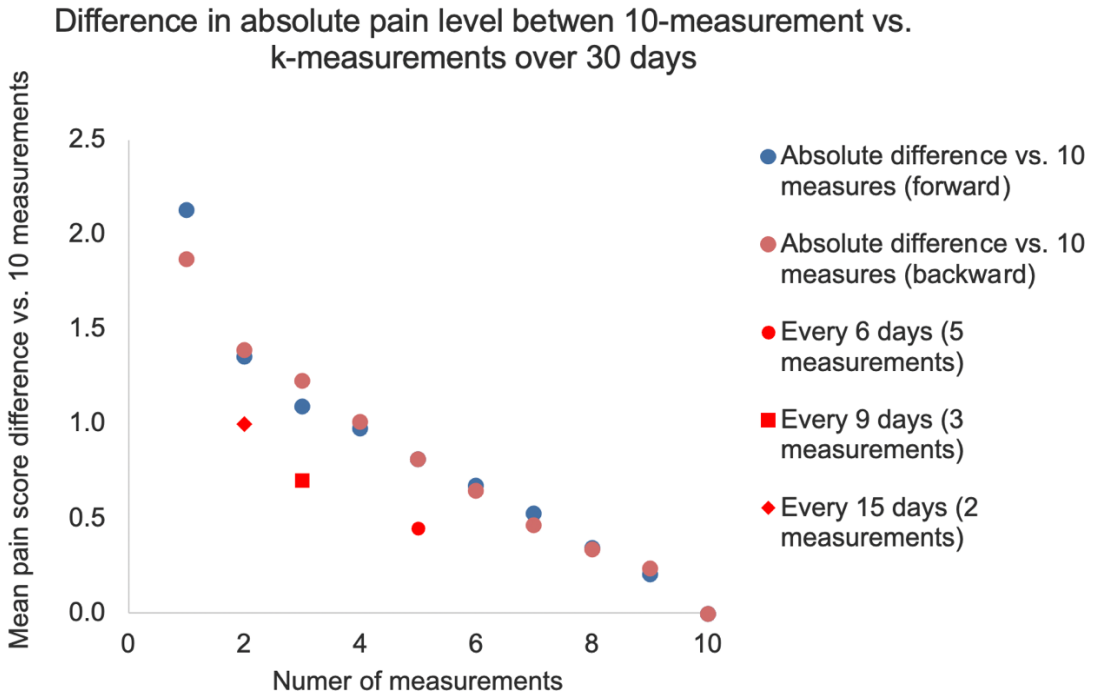
*Each line represent a patient

*Solid line is the study population mean. Dotted lines are the 95% confidence interval.

*Dotted lines are the observed mean for each trajectory class. Solid lines are the fitted line. Shades are the 95% confidence interval.

Trajectories of pain 30 days after surgery, at individual level (A), cohort-level mean (solid line, Panel B) and 95% confidence interval (dotted lines, Panel B), and by latent class (C). In Panel C, percentage values in the parenthesis indicate the mean probability of the classified patients belonging to the particular class, dotted lines are the observed mean pain level in each class, solid lines represent fitted lines, and colored band represent 95% confidence interval. The figure shows that cohort-level mean oversimplifies variable trajectories and the latent class model may offer an interpretable representation of various trajectories.

Figure 3: Representativeness of pain experience characterized via variable number of measurements



Difference in patient-level pain scale ranging 0-10 (y-axis) over 10 measurements vs. over k measurements (x-axis). Using only one measurement of pain at the beginning or end of the measurement yielded approximately 2-point differences when compared with average pain level of 10 measurements over 30 days. Blue circles are counting number of measurements from the first postoperative survey and orange circles are counting from the last survey. This difference diminished incrementally with every 15, 9, and 6 days of measurements.

Supplementary Materials

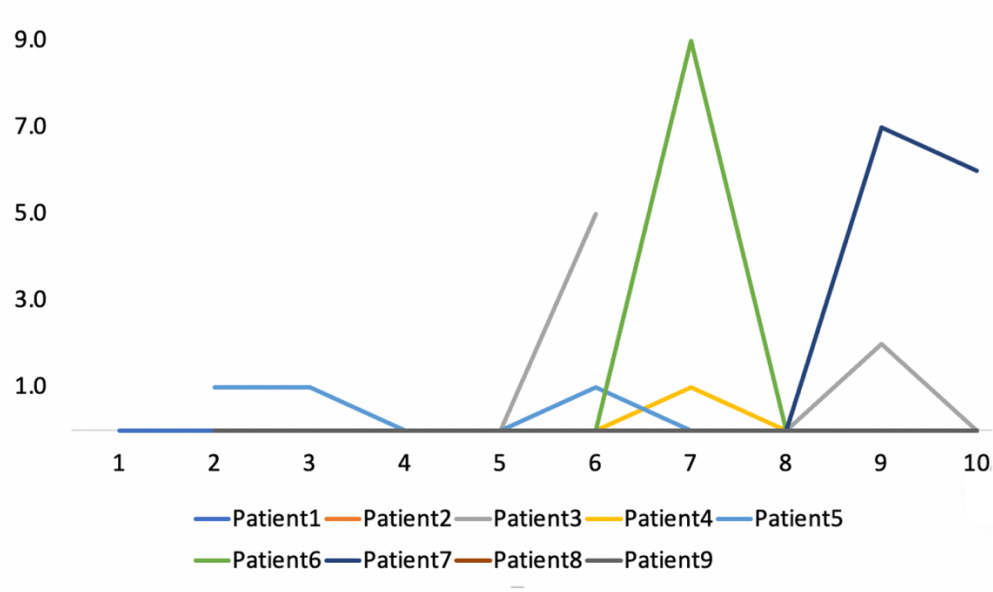
Appendix Table 1: Bayesian Information Criterion value by the number of trajectory classes

Number of trajectory class	BIC value
1	1455.46
2	1285.72
3	1279.55
4	1259.05
5	1259.31

BIC = Bayesian Information Criterion

Based on the best (lowest) BIC value, we selected 4-class model.

Appendix Figure 1: Observed pain scores of patients classified into the persistently low trajectory group



The figure shows the reported pain levels of the 9 patients who were classified to be in the persistently low pain trajectory group. Most patients in this group (6/9) scored 2 or lower across the entire follow-up duration while there were 3 patients who scored outlier values at 5th or later measurement. The mean score at each time point for the 9 patients ranged from 0.0 to 1.7, confirming that this patient group had overall low pain score during the follow-up period.

CHAPTER 3

A study design and analytical approach to measure and characterize patient-centered postoperative recovery

Introduction

Improving postoperative patient recovery is a priority. Readmission rates in the post-operative period are high. Moreover, in the United States, the expansion of episode-based payments and performance measures is increasing interest in the post-acute experience of patients[1, 2]. However, we generally lack systematically-collected information on the experience of patients in the post-acute period, as few studies rigorously collecting information using established patient-reported outcomes measures (PROMs). We have, for example, little information about the variation of the trajectories of recovery and the factors most strongly associated with better outcomes[3].

The assessment of the patient experience can provide important insights into the process of recovery that is not evident through clinical outcomes or intermittent clinical office visits. PROMs and wearable devices can provide complementary information by providing measurements of how the patient's experience and functional status change over time[4]. Current digital platforms allow us to efficiently collect PROMs and wearable-generated data at high frequencies and with little cost and burden. These automated data collection approaches may minimize the bias introduced by clinician-directed patient interviews[5]. Such a platform is highly suited to obtain repeated measures to characterize a time-dependent process such as recovery[6].

Cardiac surgery is an ideal area for the study of recovery. Many patients have good outcomes, but the limited existing evidence suggests a wide variation in the post-operative

experience of these patients[7]. However, these patients' experience has been poorly studied, as most studies of recovery simply assess deaths and complications.

Characterizing the recovery from the patient perspective is important for many reasons. First, shared decision-making and informed consent should be guided not only by the risk of mortality and complications but also by the recovery experience. Understanding variations in recovery could enable the early identification of people who are struggling and require additional attention. Recovery data from the patient perspective may enable remote monitoring after the procedure to selectively and preemptively intervene on those at high risk of poor recovery to improve outcomes. Characterization of recovery can also be used to identify patient, surgeon, procedural, and institutional factors that are associated with different patterns. With this information we can identify modifiable risk factors for poor recovery.

Thus, at this juncture, there are several notable gaps in knowledge. First, although recovery occurs over time, most studies of recovery included a small number of timepoints, and the recovery trajectory phenotypes remains poorly defined[3]. Cohort-level average of recovery trajectories is a common way of reporting[3] and can indicate how patients recover on average[7], but it obscures individual variation such as rapid early recovery, gradual recovery, or initial recovery followed by a decline. Second, we have limited understanding of how recovery trajectories vary by patient factors, operation types, center or surgeon characteristics, procedural processes, and complications, which limit opportunities to identify high risk patients preemptively and intervene.

Accordingly, our overall objective is to characterize short-term trajectories of patient recovery after cardiac surgery using PROMs and wearable data. We are conducting a prospective study to characterize trajectories of postoperative recovery in multiple domains after cardiac surgery. The specific aims of this study are to: 1) leverage a digital data platform to

collect PROM and wearable device data to bring forth the variable individual recovery trajectories, 2) describe distinct classes of recovery trajectories and clinical factors associated with the classes, and 3) to evaluate whether early postoperative recovery trajectory predicts later recovery trajectory. In addition, we will investigate optimal ways to manage missing data specific to these time-series data This study is a step toward using this approach to prospectively monitor and preemptively identify patients at risk of poor recovery and facilitate intervention to reduce the risk of adverse events. The purpose of this study protocol summary is to describes a new approach to studying recovery in order to address the knowledge gap as well as to prespecify our approach.

Methods

Design Overview

This is a prospective cohort study of patients who are undergoing valve, CABG, or aortic surgery at a tertiary center in the U.S. We chose the operations because they are the most common cardiac operations performed[8] while having different patient and operative characteristics, such as the use of deep hypothermic circulatory arrest, to potentially provide insights into the recovery pattern associated with such variations. Subgroup analysis will be conducted to evaluate whether there is a distinct patient experience by operation types. We are enrolling patients postoperatively after ICU discharge in order to ensure clinical stability, and we electronically delivering surveys directly to patients every 3 days for 30 days after hospital discharge to study patient trajectories in multiple domains characterizing recovery. The closing phone interview after 30 days, electronic medical record review, and linkage to the Society of Thoracic Surgeons database are used to confirm survival, readmission, and complications. The closing interview asks about details of readmissions if they occurred, patients' overall satisfaction with the study, and whether their experience was well captured by

the summary of their PROM data. We will apply group-based trajectory modeling to the longitudinal PROM data to identify distinct categories of recovery trajectories in a data-driven fashion. We also identify predictors of protracted recovery trajectory and evaluate whether early recovery patterns (<10 days) predict the overall trajectory (30 days) at the patient-level. The Yale Institutional Review Board approved this study (IRB # 2000025689).

Patient Population

This study began in January 2019 and is ongoing. The study is taking place at Yale-New Haven Hospital, a tertiary center in the United States, where over 1,100 cardiac surgeries are performed annually. Inclusion criteria are patients of age 18 and older who are undergoing coronary artery bypass grafting (CABG), valve replacement or repair, or aortic operations. Exclusion criteria are those who undergo heart transplant, extracorporeal membrane oxygenation (ECMO), adult congenital operations, or ventricular assist device implantation, as these patient populations tend to have a longer course of intensive care unit stay[9], precluding the timely enrollment necessary to capture immediate postoperative recovery. We also excluded those who do not own a smartphone or a tablet or those who do not speak or read English, because the digital platform for PROM data collection relies on patients responding to surveys displayed on web browser via email or text, and the surveys were written in English language. We do not allow proxy for survey response and consequently excluded patients who were not able to respond by themselves as determined by the research assistant.

In order to provide the sense of patient selection resulting from these criteria, we will compare patient characteristics of those who were approached and were and were not able to participate in the study for any reasons.

Recruitment

Recruitment takes place postoperatively after the patient has left the intensive care unit (ICU) for the step-down or floor unit (Figure 1). We chose to enroll patients postoperatively, as opposed to preoperatively, because postoperative enrollment allows for enrollment of patients who undergo surgery under non-elective settings. Recruitment after transfer from the ICU setting ensures clinical stability. A research assistant (RA) visits the patient and after confirming the patient is eligible to participate and following the description of the study procedure, obtains written informed consent (Supplementary Material S1) from all study participants. The informed consent form states that all personal information, survey response, and any medical records are confidential, will not be shared, and will be stored in an encrypted database.

We iteratively refined the enrollment process to minimize the onboarding time, which includes obtaining informed consent and signup process directed by the RA on a tablet device to enter patient name and email address or phone number and takes approximately 10-15 minutes.

PROM instrument and administration

We use 24-item quality of recovery (QoR-24) to characterize patients' postoperative recovery in various domains. The questionnaire consists of 24 items that were developed and validated in inpatient and outpatient surgical populations in terms of convergent validity with visual analogue scale, construct validity compared with length of hospital stay and sex-based difference, along with good internal consistency and test-retest reliability[10-13]. We chose QoR-24 among 5 other PROMs developed specifically to measure postoperative recovery. QoR-24 possessed many qualities advantageous for the purpose of our study, including the robust validation of psychometric property, extensive use cases in various surgical populations, ability for self-administration, and the ease of interpreting item-wise scores (Supplementary Table 1-

2). The instrument was previously adapted into a mobile format and was successfully used to administer the survey daily for 14 days[11, 12]. We added 3 items to QoR-24 to capture the self-reported time patients went to sleep, the time they awakened, and their global perception of how much they have 'recovered' in a 0-100% scale. The resulting 27-item questionnaire takes 2-4 minutes to complete, making its frequent administration feasible (Supplementary Material S2). Among the published studies in cardiac surgery, this study will have the highest number of PROM data points collected in the first postoperative month[3].

Digital data platform

We are delivering surveys on the day of enrollment and every 3 days for 30 days. This method provides detailed longitudinal data across multiple domains of recovery (Figure 2). To facilitate data organization and scheduled survey delivery, we use Hugo (Me2Health, LLC, Guilford CT, USA) a patient-centered health data sharing platform, which has a customizable survey delivery function and reminder feature to facilitate data collection. Hugo platform allows for automated delivery of surveys without researchers having to directly contact patients, which facilitates high-frequency data collection. Additionally, it imports data from connected wearable devices to facilitate centralization of patient health data. The patients retain access to their own data in a cloud-based account. Hugo does not fall under the Covered Entity that Health Insurance Portability and Accountability Act (HIPAA) regulates, but employs all the security measures that would be required by HIPAA had it been a Covered Entity.

Identifying common reasons for low response rate

Recognizing that the survey response will be incomplete for some participants, we have conducted a phone interview with the first 22 patients to learn reasons for low responses and identify strategies to minimize the barriers toward survey response for subsequent participants.

In the first 22 patients, we identified 5 with response rate of <50% and conducted recorded phone interviews. Our interview guide (Supplementary Material S3) contained questions to elucidate technical barriers, differential preferences for engagement, and or any other issues precluding survey completion. We also asked whether the length of the questionnaire or types of questions asked made it difficult to complete the survey. Two members of the research team (CB and MM) evaluated the interview recordings to identify common reasons for low response rate. This suggested the potential importance of reminder to maintain patient engagement. We modified the protocol to contact all participants approximately 10 days after enrollment. We will continue to conduct this phone interview for patients with low response rate and describe engagement and barriers to participation in the final cohort. Survey response rate and time spent to complete each survey will be reported descriptively to evaluate the degree of patient engagement. This approach likely allows us to identify patients who either did not respond or completed the survey in an unrealistically short time that may not represent a meaningful response.

Additional clinical data and adjudication of hospitalization and survival

Additionally, we are using the Society of Thoracic Surgeons (STS) Adult Cardiac Surgery Database data specifications to retrospectively collect clinically relevant data in this patient population. Pre-specified candidate predictors in this database will be used to identify clinical predictors of recovery trajectories (Table 1). The STS database contains patient demographics, comorbidities, presenting clinical status, operative details, and postoperative mortality and morbidity up to 30 days after the time of operation[14]. These data are routinely collected at Yale New Haven Hospital. At our program, 30-day mortality rates for isolated aortic valve

replacement and isolated CABG are stable around 1%, with 30-day readmission rate of about 10%, which are slightly lower than the national average.

We will determine mortality and hospital readmissions by several approaches: review of hospital records, review of cardiac surgery clinic notes, and conducting closing phone interviews with the patient or contact person previously identified.

Patient Involvement

Prior to launching the study, we interviewed 5 patients both in pre and postoperative settings to evaluate whether the frequency of survey delivery and PROM instrument were likely to adequately capture their experience of recovery. All patients agreed that the frequency of questionnaire administration and the length of the PROM instrument were reasonable and provided face validity that the questionnaire captured aspects of recovery that were important to the patients. Additionally, this article is authored with a patient (LG) who participated in the study to reflect his perspective on the study design and experience in responding to the surveys.

Sample size

The study sample target is 200 patients. Adequate sample size for studies using group-based trajectory modeling depends on the dataset's representativeness of the population of interest[15]. Therefore, the concept of statistical power traditionally used for sample size calculation does not apply to latent class analyses. We may generate a larger simulation dataset from the measured patient trajectory data to perform a split-sample testing, evaluating whether trajectories generated from the derivation sample would allow for satisfactory categorization of the testing dataset. Additionally, the study setting is scalable to increase the sample size by increasing the enrollment period, should a larger sample size become necessary.

Analytical approach – group-based trajectory modeling

The resulting dataset is a complex time-series data, with each patient having 10 data points (one every three days) at different postoperative times for each item. A practical approach to dimension reduction is group-based trajectory modeling, which is a type of latent class analysis that groups similar patient trajectories according to a number of features derived from the time-series data[16, 17]. This approach allows for dimension reduction of the complex time-series data into several distinct classes of recovery trajectories. These trajectories can be labeled according to the observed clinical phenotype of trajectories, for example 'fast recovery,' 'average recovery,' or 'protracted recovery,'. This data-driven categorization enables additional regression modeling to identify predictors of patients belonging to a certain class of recovery path.

The dataset will be classified into distinct categories of trajectories at domain level, using group-based trajectory modeling[16, 17]. Traj package on R[18] or Proc Traj package on SAS[15], performs trajectory modeling by first extracting 24 features of patient-level trajectory, selecting a subset of features that describes the overall trajectory, and identifying optimal number of classes to group the trajectories based on the longitudinal k-means method. The 24 features include range, mean change per unit time, and slope of the linear model (Table 2), which have been demonstrated to discriminate between stable-unstable, increasing-decreasing, linear-nonlinear, and monotonic-nonmonotonic patterns of trajectories[18]. K-means method partitions the time-series data into k groups such that the mean squared error distance of each data point from the assigned cluster is minimized[19]. The optimal number of clusters is determined by the minimization of Bayesian information criterion, which signifies the balance between model's complexity and the ability to describe the dataset. This process yields distinct classes of patient

trajectories in a data-driven fashion. Trajectories will be identified separately for the 5 domains and 1 global recovery measure.

With the characterization of trajectories, we will then fit multinomial logistic regression models using clinical variables outlined in Table 1, including patient demographics, comorbidity, and postoperative event such as complications and ICU readmissions, to identify predictors of patients belonging to each trajectory class. As some variables interact with each other, such as history of chronic lung disease increasing the risk of postoperative pneumonia, which likely impacts the recovery experience, we plan to stratify the cohort with and without the index complications defined by the STS (prolonged ventilation, renal failure, sternal wound infection, pneumonia, stroke, all-cause reoperation). Further analyses on interaction and mediation effects likely requires a larger sample size and are of interest in the future.

Analytical approach – missing data

Because missing data are inevitable in longitudinal PROMs, there is a need employ an appropriate handling of missing data. Multiple imputation prior to latent class analysis may yield a less biased estimate of the resulting trajectories. An alternative approach used in group-based trajectory models assumes the data are missing at random (MAR) and generates the maximum likelihood of the model parameters[20]. MAR is valid when the response attrition is independent of the group membership. However, patient attrition is oftentimes dependent on clinical characteristics and likely related to the class of trajectory itself. An extension of the model allows for modeling of attrition across trajectory groups[21], permitting dropout probability to vary as a function of covariates or observed outcomes prior to dropout and yields a more robust estimate of the probability of group membership. As such, we will perform sensitivity analysis to compare the trajectories generated via raw data vs. data preprocessed

with multiple imputation vs. trajectories generated via trajectory model accounting for response attrition.

Results

Between January and May 2019, we have enrolled 22 patients who completed the 30-day follow-up. In this cohort, median age was 58.5 years (interquartile range 53.5-67.0) and 7 (32%) were women. There were 9 (41%) mitral valve repair cases and 6 isolated or concomitant CABG (27%).

Barriers to completing surveys

Of the 22 patients enrolled, 3 (14%) did not complete any surveys, 19 (86%) completed at least 3 surveys, and 17 patients (77%) completed at least 6 of 11 delivered surveys (>50% of delivered surveys). Of the 5 patients who completed less than half of the surveys, we successfully contacted 4, and 1 could not be reached after 5 attempts. All 4 reported that the major barriers precluding survey completion were their clinical conditions: 2 described readmissions as an overwhelming event that made them feel continuing survey participation challenging, and 2 described not feeling well in general, which precluded participation. All 4 patients noted that text or email reminders might have been helpful to sustain participation. Based on these responses, we modified the protocol to contact all participants approximately 10 days after enrollment to improve engagement and resolve any patient-specific issues in completing the surveys.

Clinical outcomes

There were no deaths during follow-up. Two (9%) patients experienced at least 1 hospital readmission. Figure 2 depicts the breadth in recovery trajectories in pain, sleep, ability

to take care of own hygiene, and perception of overall recovery in five patients with complete response.

Discussion

This study will provide time-series data on short-term recovery after cardiac surgery using PROM instruments complemented by clinical records obtained via the STS database and electronic health records. This study will provide one of the highest density of postoperative PROM data in existing cardiac surgery literature[3], and it will characterize the variability in individual recovery processes with a high temporal resolution. This study will be important in closing knowledge gaps around patient-level variations in trajectories because prior studies have mostly focused on changes in PROM scores at a limited number of time points[3] or reporting group-level aggregate of longitudinal recovery data[7, 22]. Because recovery is an individual, variable, and time-dependent process, we designed our data collection and analytical approach to capture such features important to recovery.

This study has the potential to make a variety of contributions toward improving post-acute phase of care. First, we will be able to develop a preliminary nomogram of postoperative recovery for each domain and overall perception of recovery, which would be instrumental for patients and clinicians to gauge the breadth of possible recovery trajectories to facilitate informed shared decision-making. Second, identifying predictors of accelerated or protracted recovery, as classified by group-based trajectory model, may allow for individualized prediction of the postoperative recovery course to better inform the patients and family members. Third, early detection of recovery signals related to adverse events, such as mortality and readmission, may eventually facilitate preemptive intervention and focused monitoring of patients at an elevated risk for such events. Our design of the longitudinal PROM data

collection allows for incremental update of such prediction as patients progress through the phase of recovery.

There are many challenges to the successful acquisition of patient measurements during recovery: efficient administration of PROMs in a way that does not require prohibitive amount of resources, minimizing selection bias originating from barriers to survey completion, handling of missing data that inevitably occurs in PROMs, and summarizing the complex data in a way that is interpretable to surgeons and patients[23]. Additionally, the use of wearables and device data require active patient participation in periodically charging the device, wearing them correctly, and reliably syncing the device to the server for data uploads. Moreover, there is a need to provide value to the patients for providing their recovery profile, such as giving them access to their health data in a meaningful way.

The resulting data collection, analytical, and output platforms have the potential of being implemented in the clinical setting where an integration of incrementally increasing PROM and clinical data provides the near-real time estimate of individual patient risk of adverse post-operative events. Such a model may allow for triggering of preemptive clinical intervention. An output may assimilate a form of clinical dashboard within the electronic health record system, which may be monitored at a centralized location where a trained clinician reviews high-risk cases filtered by the algorithm to further evaluate whether the patient condition warrants an intervention. Together, this workflow has a tremendous potential to improve post-acute phase of care following surgery.

Lessons Learned from the initial experience

Through this first group of enrolled patients, we learned that most of the patients approached were willing to participate and consented to the study. By streamlining the

enrollment process, the enrollment time shortened from over 1 hour on the first patient to approximately 10-15 minutes for the current enrollment. The overall response rate is acceptable, with 77% of the participants completing more than half of the delivered surveys independently without any intervention by researchers. Challenging recovery course, including readmissions may have interfered with patient engagement. While this would have resulted in an underrepresentation of those with protracted recovery or with complications, our preliminary data show we were able to capture variations in the trajectories of recovery.

To sustain patient engagement through challenging recovery course, we implemented a protocol for a research assistant to call the patient around 10 days after enrollment to troubleshoot any issues and reemphasize the importance of their participation. By the protocol, research assistant making this call does not act in clinical capacity and does not provide clinical evaluation or advise, which is an important boundary for this call to not act as an intervention to alter recovery course. We believe that once the survey becomes part of clinical workflow with clinicians monitoring and responding to the PROM response, patient response rate would improve further.

We modified the enrollment protocol to reduce the enrollment time, because to some patients, the complexity and prolonged time spent for enrollment discouraged signups. Initial protocol for enrollment required patients to download an app and register. This resulted in a wide range of time spent for enrollment between 15 minutes and 90 minutes, with longer enrollment owing to technical challenges. These challenges include patients forgetting the password for app download, having to reset the password, and not having immediate email access to check account confirmation emails. Because our cardiac surgery patient population tended to be older, these technical challenges may have been pronounced. By not including the

app download and allowing for the research assistant to enroll the patient via an online form with their permission, the enrollment time shortened significantly to 10-15 minutes.

Examining the initial individual data on recovery, there were wide variations in the trajectories of recovery even among only 5 patients. The variation suggests that the instrument we used was sensitive to capturing such differences. We also noted variations in improvement over time across different domains of recovery, where overall perception of recovery seemed to have a steady improvement pattern, while pain varied between consecutive measurements in some patients.

Limitations

There are several limitations to this study. First, the single-center tertiary care setting limits the sample size and applicability of the findings to patients cared for in different settings. A multi-center study following the current study would address this limitation and evaluate whether the findings at our center are comparable to findings in other centers. Additionally, group-based trajectory modeling will classify patients into distinct trajectories based on similar recovery patterns, and this analytical approach may allow for generalization of the variations in the trajectories as long as our sample represents the breadth of the possible variation in recovery.

Another limitation is the exclusion of patients who cannot participate for various reasons. The use of digital platform is advantageous in reducing the resource intensity for data collection, but leads to exclusion of patients who do not own mobile devices, which likely affects older patients disproportionately. As the number of adults using mobile devices is increasing[24], we believe this will become less of a limitation over time. Initiating this study now despite this limitation is important to establish a platform that may become the standard of

postoperative care when the vast majority of patient population own digital devices in a predictably near future. Those who cannot participate due to lack of interest or technological barrier represent an important population that may be distinct in characteristics and risk profiles. While acknowledging the selection bias originating from this inclusion threshold, we believe there is a need to initiate collection of patient-centered outcome measures in the proposed approach, in order to further engage hospitals and programs for a broader implementation of this approach in the context of extremely limited evidence base. We plan on minimizing the non-participation for the lack of interest by intermittent phone check-ins to sustain interests and identify barriers to inform strategies to increase engagement. While recognizing that clinical implemenation of this protocol would preclude the use of incentives, in following studies, we may consider other forms of incentives to participate, if this population is indeed distinct and large in proportion. Additionally, when the PROM data are integrated into routine clinical care, patient engagement will likely increase substantially because they will be more inspired to share these data if they are used by their clinicians.

Finally, postoperative enrollment and retrospective assessment of preoperative health status, as opposed to preoperative enrollment, may introduce recall bias. We decided on postoperative enrollment, because preoperative enrollment precluded standardized enrollment of patients operated on under non-elective settings. Given the retrospective assessment of baseline health status takes place on the first postoperative survey, we believe the recall bias is minimized owing to the temporal proximity.

Latent class analysis to uncover clinical phenotypes

In precision medicine, a common question for researchers is whether patients can be classified with others who have similar risks and treatment responses. Such groupings can assist

in predicting risk and matching patients with appropriate treatment strategies. The challenge is that it is often not easy to identify meaningful clusters of people with the observable data.

Latent class analysis (LCA) is a common explanatory modeling technique that allows researchers to identify groups of people that have similar characteristics that can include demographics, clinical characteristics, treatments, comorbidities, and outcomes(1). The term latent derives from the fact that the classes are not directly observable. LCA estimates the probability of each participant being a member of each latent class (2).

In the November 17, 2019 issue of *JAMA Cardiology*, Patel, *et al*. (3) used group-based trajectory modeling (GBTM), a type of LCA, and identified five distinct patterns of change in participant urine albumin-to-creatinine ratio (UACR) observed over 20 years. These 5 classes were independently associated with adverse changes in cardiac structure and ventricular function (3). Notably, participants belonging to the identified trajectory classes could not be distinguished by the baseline UACR alone, highlighting the value of this technique. This Guide to Statistics and Methods article describes LCA, its potential application, and limitations.

What is latent class analysis?

LCA is a statistical technique that identifies groups defined by specific combinations of observed variables (2). LCA assigns each participant a probability of being in each subgroup based on maximum likelihood estimation. Then, each participant is assigned to the group to which they have the highest probability of belonging. In GBTM, the trajectories' shape can be a straight or curvilinear form; shapes are based on the maximum likelihood estimation. Selecting the number of groups requires manual reconciliation of the trajectories' shape, the minimum number of participants assigned to a trajectory, and measures indicating how well the model fits, such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) (6).

Although there is no single criterion to select the number and shape of classes or trajectories, the number of classes that yield the best fit to the observed data, the highest average probability of group membership, and the fewest poor fitting participants (i.e. those with a highest probability of group membership <0.7) is chosen (7). Therefore, reporting the decisions and rationales behind this process of manual reconciliation is crucial.

Why are latent class analyses used?

LCA is useful when the patterns that constitute distinct clusters or classes are difficult to discern with traditional methods. For example, the clinical heterogeneity within the broad definition of sepsis has made it difficult to determine whether there are patient subgroups that respond more favorably to one treatment or another. Investigators have used LCA as a confirmatory analysis (4) to reproduce novel phenotypes of sepsis identified by another clustering technique called consensus k-means clustering. The investigators identified clinical phenotypes of sepsis with differential treatment responses, based on a combination of hemodynamics, laboratory, and end-organ functional parameters.

LCA can also capture groups of participant preferences that depend on a complex intersection of options. For example, LCA was used recently to group personal preferences for bariatric surgery resulting in three subgroups relating to: concerns with costs, benefit-focused and procedure-focused (5). An advantage of this approach is that the grouping originates from the data; thus, the categories are not predefined and thus not limited by current conceptual frameworks.

The LCA methods include longitudinal approaches, in which participant-level trajectories of an outcome can be classified into groupings. These are called GBTM or latent class growth analysis (6) and identify underlying subgroups that would have been masked if only a single

regression line was estimated, as done in the majority of longitudinal analyses. For example, Wu, et al. (1) found five trajectories of overall cardiovascular health over four years that were independently associated with subsequent risk of incident cardiovascular disease.

Limitations of the latent class analysis

Several limitations of LCA merit consideration. First, although grouping based on latent class facilitates data presentation and interpretation, participants do not actually belong to a single group. The class membership for each participant is assigned based on the highest probability of belonging to one of the latent classes. That is, some participants have similar probabilities of belonging to multiple groups (i.e. probabilities of 0.5, 0.49 and 0.01 for classes A, B and C, respectively); however, the group membershio is assigned based on the highest probability (7). Therefore, it is critical to examine the participants for whom the highest probability of belonging to a single class is poor (<0.7) and provide descriptions of such participants (7).

Second, the number of classes is derived from the cohort considering the model fit and complexity and that the number is not fixed. LCA applied to a larger cohort or a cohort with more observed characteristics may yield a different number of classes with different patterns. Therefore, reporting validation and reproducibility of the latent class is important. To validate the latent classes, researchers may perform cross-validation. Reproducibility should be tested by using a different source data (4) to test whether the identified groupings can be reproduced, although it is often difficult to find an independent dataset of similar cohort that contains variables comparable to the original dataset.

Third, for GBTM applied to longitudinal data, participants assigned to a class may vary around the estimated trajectory. For example, a participant may assimilate a rapid recovery

trajectory initially, then experience a catastrophic event such as stroke, and no longer follow the initial trajectory afterward (8). In such cases, the probability of membership is unlikely to be high for a single class. Furthermore, GBTM is only able to model trajectories in polynomial functions and is not equipped to model trajectories that do not conform to other shapes, such as a cyclical trajectory.

How were latent class analyses used?

In the study by Patel, *et al*.(3), authors used GBTM to categorize participants (n=2,647) into five distinct classes of trajectories of UACR (Figure 1 in the article by Patel, et al.) over the course of 20 years in young adults. UACR measurements were recorded prospectively at five timepoints (10, 15, 20, 25, and 30 years since the enrollment). GBTM identified distinct categories of trajectories, that were not identifiable from the baseline UACR measurement alone. The authors labeled the trajectories, with the 'high-increasing' group (1.6% of participants) showing a clinically concerning pattern of persistently high UACR level that continued to increase over the study period. Their linear models showed that trajectories of worse UACR were associated with greater risk of adverse cardiac structural alterations and ventricular systolic and diastolic functions measured at year 30, the end of the study period.

How should the latent class analysis be interpreted?

Using GBTM, Patel, *et al*.(3), were able to reduce the complex longitudinal UACR data to an interpretable number of trajectory types that prognosticated adverse cardiac functions and structural alterations. They concluded that such trajectory-based categorization may help with early identification of those at risk of subclinical cardiovascular disease. However, identifying the group expected to have the worst trajectory before the completion of follow-up remains a

challenge. As expected in a longitudinal study, there were participant attritions over time (71% of surviving participants completing the last follow-up). Additionally, the analytical method required each participant to have at least three longitudinal measurements to be included in the analysis, leading to an additional exclusion. The excluded cohort differed from the retained cohort in certain demographic and comorbidity characteristics. As these exclusions occurred based on follow-up information, it remains unknown whether the trajectory classes are generalizable to the initial cohort prior to such exclusions, which is the cohort for which clinicians are interested in prognosticating the risk. There are methods for modeling loss over follow-up using GBTM (6), which were not applied. As the authors modeled the trajectories of >2600 participants, examination of a split sample may have reinforced the reproducibility of the latent classes. Regardless, the general conclusion that the dynamic changes of UACR may be associated with a later adverse cardiac remodeling is supported by their approach.

Conclusion

This study will generate highly granular, longitudinal PROM data to characterize individual trajectories of patient recovery after cardiac surgery. Digital data sharing platforms promise to minimize the patient and researcher burden in administering and completing PROMs, allowing for characterization of granular progression of patients' state of health over time in the postoperative period. Implementation of such study is complex but feasible, and it will serve as an important platform to facilitate clinical use of PROM data to improve the overall patient recovery. Latent class analysis may provide insights into the underlying heterogeneity of recovery trajectories that have not been available via conventional, investigator-driven grouping of patients.

References

1. Wadhera RK, Yeh RW and Joynt Maddox KE. The Rise and Fall of Mandatory Cardiac Bundled Payments. *JAMA*. 2018;319:335-336.
2. Khera R, Dharmarajan K, Wang Y, Lin Z, Bernheim SM, Normand ST and Krumholz HM. Association of the Hospital Readmissions Reduction Program With Mortality During and After Hospitalization for Acute Myocardial Infarction, Heart Failure, and Pneumonia. *JAMA Netw Open*. 2018;1:e182777.
3. Mori M, Angraal S, Chaudhry SI, Suter LG, Geirsson A, Wallach JD and Krumholz HM. Characterizing Patient-Centered Postoperative Recovery After Adult Cardiac Surgery: A Systematic Review. *J Am Heart Assoc*. 2019;8:e013546.
4. Gill TM, Gahbauer EA, Han L and Allore HG. Trajectories of disability in the last year of life. *N Engl J Med*. 2010;362:1173-80.
5. Pakhomov SV, Jacobsen SJ, Chute CG and Roger VL. Agreement between patient-reported symptoms and their documentation in the medical record. *Am J Manag Care*. 2008;14:530-9.
6. Moore FD. Getting well: the biology of surgical convalescence. *Ann N Y Acad Sci*. 1958;73:387-400.
7. Diab MS, Bilkhu R, Soppa G, Edsell M, Fletcher N, Heiberg J, Royse C and Jahangiri M. The influence of prolonged intensive care stay on quality of life, recovery, and clinical outcomes following cardiac surgery: A prospective cohort study. *J Thorac Cardiovasc Surg*. 2018;156:1906-1915.e3.
8. Thourani VH, Badhwar V, Shahian DM, O'Brien S, Kitahara H, Vemulapalli S, Brennan JM, Habib RH, Fernandez F, D'Agostino RS, Lobdell K, Rankin JS, Gammie JS, Higgins R, Sabik J, Schwann TA and Jacobs JP. The Society of Thoracic Surgeons Adult Cardiac Surgery Database: 2019 Update on Research. *Ann Thorac Surg*. 2019;108:334-342.
9. Blanche C, Blanche DA, Kearney B, Sandhu M, Czer LS, Kamlot A, Hickey A and Trento A. Heart transplantation in patients seventy years of age and older: A comparative analysis of outcome. *J Thorac Cardiovasc Surg*. 2001;121:532-41.
10. Dahlberg K, Jaensson M, Eriksson M and Nilsson U. Evaluation of the Swedish Web-Version of Quality of Recovery (SwQoR): Secondary Step in the Development of a Mobile Phone App to Measure Postoperative Recovery. *JMIR research protocols*. 2016;5:e192.
11. Jaensson M, Dahlberg K, Eriksson M and Nilsson U. Evaluation of postoperative recovery in day surgery patients using a mobile phone application: a multicentre randomized trial. *Br J Anaesth*. 2017;119:1030-1038.
12. Halleberg Nyman M, Nilsson U, Dahlberg K and Jaensson M. Association Between Functional Health Literacy and Postoperative Recovery, Health Care Contacts, and Health-Related Quality of Life Among Patients Undergoing Day Surgery: Secondary Analysis of a Randomized Clinical Trial. *JAMA Surg*. 2018;153:738-745.
13. Myles PS, Weitkamp B, Jones K, Melick J and Hensen S. Validity and reliability of a postoperative quality of recovery score: the QoR-40. *Br J Anaesth*. 2000;84:11-5.
14. O'Brien SM, Feng L, He X, Xian Y, Jacobs JP, Badhwar V, Kurlansky PA, Furnary AP, Cleveland JC, Lobdell KW, Vassileva C, Wyler von Ballmoos MC, Thourani VH, Rankin JS, Edgerton JR, D'Agostino RS, Desai ND, Edwards FH and Shahian DM. The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 2-Statistical Methods and Results. *Ann Thorac Surg*. 2018;105:1419-1428.
15. Loughran T and Nagin DS. Finite Sample Effects in Group-Based Trajectory Models. *Sociological Methods & Research*. 2006;35:250-278.

16. Nagin DS and Odgers CL. Group-based trajectory modeling in clinical research. *Annu Rev Clin Psychol*. 2010;6:109-38.
17. Savage SA, Sumislawski JJ, Bell TM and Zarzaur BL. Utilizing Group-based Trajectory Modeling to Understand Patterns of Hemorrhage and Resuscitation. *Ann Surg*. 2016;264:1135-1141.
18. Leffondré K, Abrahamowicz M, Regeasse A, Hawker GA, Badley EM, McCusker J and Belzile E. Statistical measures were proposed for identifying longitudinal patterns of change in quantitative health indicators. *J Clin Epidemiol*. 2004;57:1049-62.
19. Hartigan JA and Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*. 1979;28:100--108.
20. Jones BL, Nagin DS and Roeder K. A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories. *Sociological Methods & Research*. 2001;29:374-393.
21. Haviland AM, Jones BL and Nagin DS. Group-based Trajectory Modeling Extended to Account for Nonrandom Participant Attrition. *Sociological Methods & Research*. 2011;40:367-390.
22. Petersen J, Vettorazzi E, Winter L, Schmied W, Kindermann I and Schäfers HJ. Physical and mental recovery after conventional aortic valve surgery. *J Thorac Cardiovasc Surg*. 2016;152:1549-1556.e2.
23. Calvert M, Kyte D, Price G, Valderas JM and Hjollund NH. Maximising the impact of patient reported outcome assessment for patients and society. *BMJ*. 2019;364:k5267.
24. Anderson M, Perrin A. Tech Adoption Climbs Among Older Adults. https://www.pewinternet.org/wp-content/uploads/sites/9/2017/05/PI_2017.05.17_Older-Americans-Tech_FINAL.pdf. Published 2017. Accessed March 30, 2019.

Tables and Figures

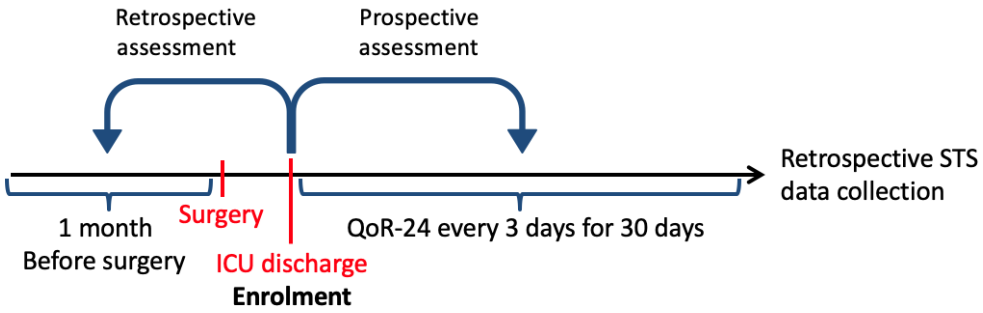
Table 1: Candidate predictors of recovery trajectory

Demographic	Comorbidity	Operative factors	Postoperative factors
Age	Diabetes	Cardiopulmonary bypass time	Length of ICU stay
Sex	Prior stroke	Cross clamp time	Length of hospital stay
Race	Congestive heart failure	Operation type	Surgical site infection
Insurance status	Chronic kidney disease	Non-elective status	Prolonged ventilation
BMI	Dialysis	Transfusion requirement	Transfusion requirement
	Prior MI	Minimally invasive approach	Stroke
	Prior cardiac surgery		Reoperation for any reasons
	Ejection fraction		Death
	Arrhythmias		Readmission
	Prior PCI		Pneumonia
	Cardiogenic shock		
	Hypertension		
	Dyslipidemia		
	Smoking status		
	Chronic lung disease		
	Endocarditis		
	Pneumonia		
	Peripheral artery disease		
	Immunocompromised		
	Mechanical circulatory support use		
	Valvular disease severity		

Table 2: 24 features of trajectory used in group-based trajectory model

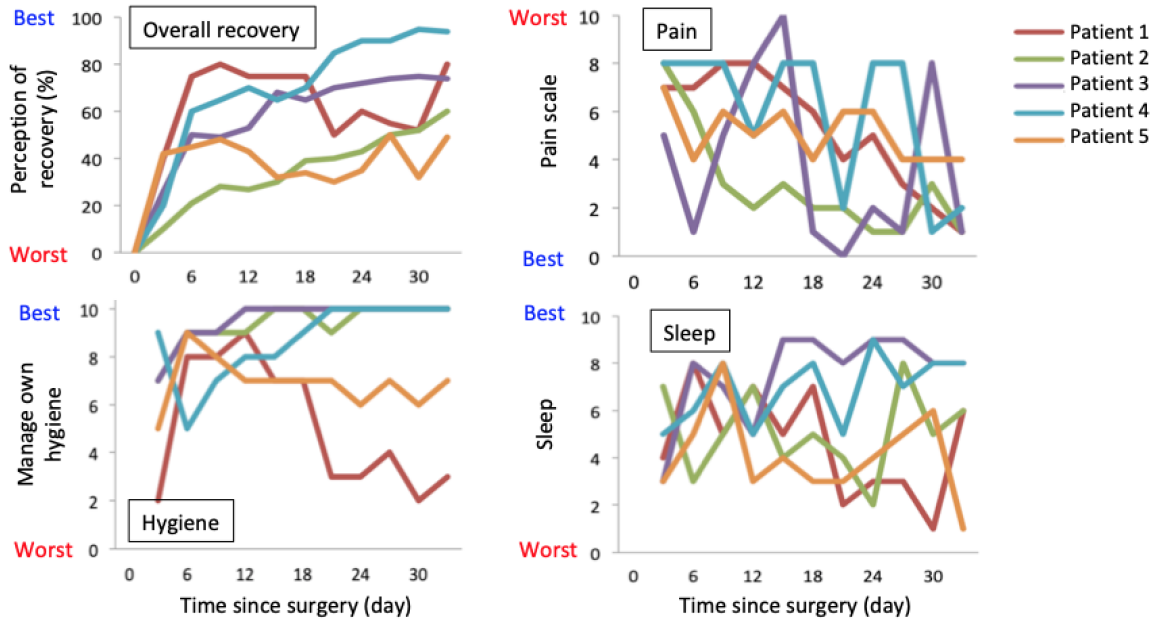
N	Features
1	Range
2	Mean-over-time
3	Standard deviation (SD)
4	Coefficient of variation (CV)
5	Change
6	Mean change per unit time
7	Change relative to the first score
8	Change relative to the mean over time
9	Slope of the linear model
10	Proportion of variance explained by the linear model
11	Maximum of the first differences
12	SD of the first differences
13	SD of the first differences per time unit
14	Mean of the absolute first differences
15	Maximum of the absolute first differences
16	Ratio of the maximum absolute difference to the mean-over-time
17	Ratio of the maximum absolute first difference to the slope
18	Ratio of the SD of the first differences to the slope
19	Mean of the second differences
20	Mean of the absolute second differences
21	Maximum of the absolute second differences
22	Ration of the maximum absolute second difference to the mean-over-time
23	Ratio of the maximum absolute second difference to mean absolute first difference
24	Ratio of the mean absolute second difference to the mean absolute first difference

Figure 1: Timing of patient enrollment and PROM administration



The figure shows the timing of patient enrollment and PROM administration over the clinical course. Baseline function is assessed by retrospectively asking the patient about their state of health during 1 month prior to the operation. 24-item Quality of Recovery questionnaire is administered every 3 days for 30 days following discharge from the intensive care unit.

Figure 2: Sample trajectories of recovery in 5 patients



The figures display trajectories of recovery in different domains in 5 patients. Each color corresponds to the same patient. Overall recovery is the patient's perception of overall recovery in 0 to 100% scale. Pain in surgical site is reported in 0 to 10 point scale, with 10 representing the worst pain. Being able to take care of own hygiene is reported in 0 to 10 point scale, with 10 representing complete independence in managing own hygiene. Patient's perception of sleep quality is reported in 0 to 10 point scale, with 10 being the best sleep.

CHAPTER 4

Examining the impact of adding intraoperative variables in predicting postoperative outcomes

Introduction

The Society of Thoracic Surgeons (STS) Adult Cardiac Surgery Database (ACSD) risk models[1, 2] are based on logistic regression and only incorporate information available before the operation to inform preoperative decision making and counseling[2]. Intraoperative events may influence the risk of postoperative outcomes[3]. Existing risk models in clinical use, such as the STS models and EuroSCORE, do not consider intraoperative information, although such data could improve postoperative prediction and patient care. A possible benefit of such dynamic update of predicted risk includes quantitatively recalibrating patient and provider expectations based on intraoperative events, which may modify decision thresholds to pursue diagnostic tests such as head scans for questionable neurologic deficit or early preparation of dialysis catheter access for those with renal failure risk that increased because of intraoperative events.

Several studies have demonstrated the value of risk models that update risk estimates as more data are generated[4-6]. In a digital era, this updating could be automated and made available to support decisions at the bedside. However, whether adding intraoperative variables improves prediction, which variables are most important, and which analytic methods yield the more accurate predictions remain unknown. Accordingly, using the national STS ACSD dataset for coronary artery bypass graft (CABG), we sought to determine whether adding and which intraoperative variables into the existing STS preoperative model would improve the predictive performance of the model.

To address this aim, we used the STS database, which includes approximately 100 intraoperative variables related to coronary artery bypass graft (CABG). We also employed machine learning approaches in addition to logistic regression. Machine learning techniques based on tree-based models, such as gradient descent boosting, are suited to identify complex relationships in high-dimensional data. Although researchers have tested machine learning

approaches to estimate risk for patients undergoing percutaneous coronary interventions[7, 8], such methods have not been evaluated extensively in cardiac surgical risk modeling.

Methods

Cohort definition and data source

We included adult patients who underwent isolated CABG from July 2014 to December 2016 in the U.S. centers participating in the STS reporting. We used the STS ACSD data definition version 2.81. We excluded concomitant cases defined at those undergoing any concomitant cardiac operations except for pacemaker implantation, arrhythmia correction surgeries, or left atrial appendage ligation or occlusion. The criteria yielded 378,834 operations. We excluded 102 cases missing gender and 160 with intraoperative death. These exclusions yielded 378,572 operations performed by 2,730 surgeons in 1,083 centers.

The STS ACSD includes >90% of the cardiac surgery centers in the United States[9]. Clinical sites enter data using uniform STS definitions for patient characteristics and outcomes. The quality of the data has been rigorously validated by comparison with independent national and local datasets[10]. The database is deidentified, and the Participant User File (PUF) Research Program Committee of the STS Workforce and the Yale University Human Investigation Committee approved this study.

Outcomes

We studied 7 postoperative outcomes using standard STS ACSD definitions: operative mortality, defined as postoperative death from any cause either in-hospital or in discharged patients, within 30 days of the index operation (with the exclusion of intraoperative deaths), prolonged ventilation, defined as mechanical ventilation requirement >24 hours postoperatively, pneumonia, permanent stroke, defined as a neurologic deficit of abrupt onset

caused by a disturbance in blood supply to the brain that did not resolve within 24 hours, reoperation for any reasons during the index hospitalization, deep sternal wound infection (DSWI), and renal failure defined as new dialysis requirement, increase in serum creatinine three times greater than the baseline and absolute rise of greater than 0.5mg/dl, or creatinine level >4mg/dL. We fitted models predicting postoperative renal failure on the data, excluding those with preoperative renal failure with or without dialysis need as done in previous works[1, 2]. Further specifications of the STS ACSD data definitions are available online[11]. Missing data were rare (<2% for all variables, except for ejection fraction, which was missing in 3%). Missing data were handled as described in the STS ACSD risk model specifications[1]. Briefly, missing data on categorical predictor variables were imputed to the lowest risk value, and missing data on continuous covariates were imputed to the conditional mean. Missing ejection fraction values were set to the mean values conditioned on congestive heart failure status and sex, and body surface area was conditioned on sex.

Candidate variables

We chose candidate preoperative variables based on the STS ACSD risk model for isolated CABG[12]. The variables were processed per the description for the STS ACSD model, which included splining of continuous variables such as age and creatinine, and combining categorical variables describing related disease states, such as congestive heart failure and New York Heart Association class variables[12]. Candidate intraoperative variables were all variables generated during the operation in the STS ACSD version 2.81. We did not employ any specific feature engineering because there are no established standards, unlike the preoperative variables. General categories of variables are summarized in Table 1, and the categories included operative approach, laboratory values, temperature measurements, transfusions,

transesophageal echocardiogram results, case duration, cardioplegia strategy, and prophylactic antibiotics use. Cardiopulmonary bypass time and cross-clamp time were excluded because we included off-pump cases. Total operative time was available for all cases and was used instead as a measure of case length. Categorical variables that would be missing in off-pump cases, such as cardioplegia-related variables, were included but missingness was treated as a feature.

Model development and validation techniques

For each of the 7 outcomes, we developed 12 models with different combinations of starting variable sets, variable selection methods, and relationship modeling for a total of 84 models (Figure 1). We used two starting variables sets: 1) preoperative variables only (same as the existing STS models), 2) intraoperative variables only, and 3) pre + intraoperative variables. Preoperative variables comprised 47 fields that were preprocessed using the method used in the previous STS ACSD risk models. Intraoperative variables consisted of 96 fields without specific preprocessing. The variables in each category are summarized in Table 1. We used two relationship modeling approaches: 1) logistic regression and 2) gradient descent boosting using XGBoost package[13]. XGBoost is a machine learning algorithm that makes a prediction based on a series of decision trees, with a highly efficient tree boosting algorithm with improved performance over other tree-based approaches in various settings[7, 13]. Its additional appeal is the ability to rank predictive variables on the order of importance to facilitate clinical interpretation of the model. We chose to use both XGBoost and logistic regression under the hypothesis that the XGBoost algorithm may yield better model performances given the increasingly large number of variables. Logistic regression is the approach that the STS ACSD risk models use. Therefore, logistic regression models were developed as a reference model

against which the performance of XGBoost models was evaluated. We used two variable selection methods: 1) no external variable selection and 2) external variable selection by support vector classifier. XGBoost algorithm has an internal variable selection process, and we hereafter refer to models with 'no variable selection' as those without prior variable selection using support vector classifier. Parameters of XGBoost models (number of trees, learning rate, and depth of trees) were tuned via internal cross-validation for each outcome to optimize c-statistics of each model, with final parameters outlined in Supplemental Table S1. For each model, we split the dataset randomly into 70% training and 30% testing dataset. This was iterated 20 times to yield 20 estimates for model performance metrics in predicting the outcomes for internal validation. We chose the number of iterations to be 20 after observing that increasing the iterations further did not change the mean or the confidence interval for operative mortality. The random sampling of the split was stratified to ensure adequate sampling of rare events. We reported means and 95% confidence intervals of 20 iterations for each metric.

Performance metrics

We evaluated the model performance for the testing dataset in each model using c-statistics, the area under the precision-recall curve (AUPRC), Brier score, resolution, and reliability. C-statistics (AUROC) characterized model discrimination and ranged between 0 to 1, with a higher value corresponding to better discrimination[14]. AUROC is the proportion of the times patients with an event were accurately classified to have a higher probability of event within all possible pairs of patients with and without an event[14]. Because AUROC can provide a misleadingly optimistic view of the model performance when classifying event of low incidences, we also evaluated AUPRC, which relates positive predictive value (also known as

precision) and sensitivity (also known as recall), and is less susceptible to unbalanced nature of datasets[15]. Therefore, AUPRC complements AUROC to characterize the discriminatory ability when the outcome of interest is rare, as it can uncover potentially faulty model performance in the precision-recall space by penalizing a high false-negative rate. Brier score is the mean squared error (MSE) of predicted probability of an event (ranges 0 to 1) and observed event (binary 0 or 1), with lower values corresponding to higher accuracy of the prediction[16]. Performance metrics were compared to their means to report whether one is numerically higher or lower, with the corresponding interpretation of better or worse. We did not evaluate the statistical significance of the difference because arbitrarily increasing the number of resampling iterations would drive the comparisons toward statistically significant differences.

We also assessed calibration using reliability measure, defined as the sum of MSE between the predicted probability and observed rate at each decile, with lower values corresponding to better calibration[17]. Reliability is more sensitive in capturing deviations of the predicted risks from the true rates than the calibration slope does. Resolution is the MSE between the deciles of predicted risks and the event rate of the entire cohort. Therefore, higher values of resolution indicate prediction across greater distances from the observed event rate and indicate models with better performance[18]. We also showed a continuous calibration plot showing model calibration for a wide range of risks using cubic spline smoothers. In contrast to the commonly used calibration plots with decile-based risk stratification, a continuous calibration plot offers an estimation of calibration in a continuum of predicted risk[19].

Clinical interpretability

To provide clinically interpretable information beyond model performance metrics, we evaluated how many cases were re-stratified according to the predicted risk generated by the

base model (logistic regression with preoperative variables without variable selection) and the model with the best performance. This step was performed on a randomly selected split of the data without multiple iterations. Risk strata were defined a priori based on authors' consensus as clinically relevant cutoffs. We applied the same threshold for outcomes with similar incidences (mortality, renal failure, and stroke as <1%, 1-3%, 3-5%, 5-10%, and >10%). We reported proportions of cases with underestimated or overestimated risks compared to the true event incidence for the base and best models. We also made this comparison between pre+intraoperative variable models fitted with logistic regression and XGBoost. We elected not to use the net reclassification index for its susceptibility to yield false-positive results when using large datasets[20].

To understand intraoperative variables that may be related to changing patients' predicted risk, we fitted logistic regression using both pre and intraoperative variables over the entire dataset to estimate coefficients, odds ratio, and 95% confidence interval of each variable in its relationship to operative mortality. For the XGBoost model, we used the 'feature_importances_' function, a built-in function in the XGBoost package, to rank the input variables in the order of importance in fitting the particular model. Additionally, we evaluated two patients whose predicted probabilies of mortality were discrepant between logistic regression and XGBoost approaches to gain further insights into how different phenotypes are handled by each algorithm.

Data preprocessing and statistical analysis were implemented with Python (version 2.7) and the open-source packages available in Scikit-Learn[21]. Three authors (MM, TJD, and CH) had access to the data, did the coding, and take responsibility for the analyses. The final code was reviewed independently by one author (AC) for quality assurance.

Results

Among the 378,572 hospitalizations for isolated CABG, the mean (standard deviation [SD]) patient age was 65.3 (10.2) years. Women comprised 93,425 (24.7%) of the cohort. Operative mortality, excluding intraoperative death, occurred in 1.9%. Permanent stroke occurred in 1.3%, renal failure in 2.1%, prolonged ventilation in 8.1%, reoperation for any reasons in 3.5%, DSWI in 0.3%. The composite event rate of the above adverse events was 12.1%. (Table 2). Intraoperative variables are summarized in Supplementary Table 3.

Preoperative, Intraoperative, Pre+Intraoperative variables

Between models using only preoperative or intraoperative variables, models for all outcomes, except for reoperation, had better AUROC values with a preoperative variable set alone than those using only intraoperative variable sets. In all outcomes, models using pre+intraoperative variables had better AUROC than respective models using either intraoperative or preoperative variable sets alone. This relationship also held for Brier score and AUPRC in all outcomes, except for DSWI, in which there were no substantial differences in Brier score between 3 different variable sets (Figure 2). Other performance metrics are summarized in Supplementary Table 3.

Logistic regression vs. XGBoost

Among the models using pre+intraoperative variables, XGBoost without prior variable selection had the best AUROC, Brier score, and AUPRC values in 4 of the 7 outcomes (mortality, renal failure, prolonged ventilation, and composite) compared with logistic regression models with or without variable selection. For DSWI, the logistic regression model with variable selection had the best performance in all 3 metrics. For reoperation and stroke, the model with the best performance varied across the metrics: discrimination (AUROC and AUPRC) was better in logistic regression models while calibration (Brier score) was better in the XGBoost model (Figure 2 for mortality and renal failure, Supplementary Table S4 for other outcomes). The

supplementary text describes two patients' phenotypes of which the predicted risk of mortality differed largely between the XGBoost and logistic regression model.

Calibration plot

Continuous calibration plot demonstrated that for all outcomes, the logistic regression model tended to underestimate the patient risk at extremely high risks compared with the observed event rate (Figure 3). Calibration plots including the confidence band and outcomes other than mortality and renal failure are summarized in Supplementary Figure 2 and 3. Calibration was generally good in the risk range, where the majority of patients resided. For all combinations of outcomes and variable sets (preoperative only or pre+intraoperative variables), XGBoost models showed better calibration across broader range of risks compared with the logistic regression model.

Risk re-stratification

The shift table of risk strata was created to compare the mortality risk stratification by the baseline model (preoperative variable set with logistic regression without variable selection) and the model that performed optimally (preoperative and intraoperative variables with logistic regression without variable selection). For mortality, this showed that baseline model underestimated the risk in 11,114 patients (9.8%) and overestimated 12,005 patients (10.6%). In contrast, the best model underestimated the risk in 7,218 patients (6.4%) and overestimated 0 patients (0%) (Figure 4). Comparing models fitted over pre+intraoperative variables, logistic regression without variable selection underestimated the risk 7,137 patients (6.3%) and overestimated in 3,566 patients (3.1%), while XGBoost without variable selection underestimated the risk in 4,263 patients (3.8%) and overestimated in 1,886 patients (1.7%) (Figure 5). Therefore, using the same set of predictors for mortality, the XGBoost model yielded 54% fewer misclassifications in risk compared with the logistic regression model. For renal

failure, using pre+intraoperative predictors, the XGBoost model yielded 112% fewer misclassifications in risk than the logistic regression model (Supplementary Table 2).

Examples of patients with large discrepancies between logistic regression and XGBoost models are the following: A 42-year-old man with minimal comorbidity who underwent emergent 3-vessel CABG received 23 units of red blood cell and 16 units of plasma. This patient had a 44% and 6% chance of death predicted by XGBoost and logistic regression, respectively. A 70-year-old man who underwent salvage 5-vessel CABG with intraoperative IABP placement. XGBoost and logistic regression models predicted the risk of this patient's mortality to be 30% and 10%, respectively. In a case where logistic regression had a significantly higher predicted mortality probability, a 52-year-old woman with minimal comorbidity undergoing elective 5-vessel CABG required IABP during the operation and had the lowest body temperature of 27°C. This patient had predicted mortality of 6% by XGBoost and 20% by logistic regression.

Intraoperative variables associated with the outcomes

Fitting logistic regression model for operative mortality using pre+intraoperative variable set, we identified intraoperative variables that improved the prediction of operative mortality. These included undergoing full sternotomy, intraoperative intraaortic balloon pump (IABP) use, lack of left internal mammary artery use, number of distal anastomosis performed, appropriate type and timing of antibiotics use, lowest body temperature, highest glucose and lowest hemoglobin level, cardioplegia route and type, blood product use, and postoperative residual tricuspid regurgitation and ejection fraction, and total case time (Table 3). Intraoperative variables identified as important features in the XGB model were similar, including: intraoperative IABP use, blood product use, timing and redosing of antibiotics, lowest body temperature, highest glucose and lowest hemoglobin level, cardioplegia route, postoperative valvular insufficiency, incision approach, and total case time. The relative

importance score for the XGB model was the highest for plasma transfusion (score of 292), followed by any transfusion (220), intraoperative IABP use (182), red cell transfusion (123), and lack of internal mammary artery use (50) (Table 4).

Discussion

Using the national STS ACSD for isolated CABG, we demonstrated that including intraoperative variables provided a better model predicting postoperative events across all outcomes, although the gain was small in some outcomes. The findings were consistent with our two analytic approaches, though the XGBoost algorithm improved the model performance slightly more than logistic regression. This study highlights the potential value of including interoperative variables and using machine learning approaches to predict the risk of adverse events after surgery.

Our study adds to the current literature in several ways. First, the existing STS ACSD models only use variables that are available before the operation, which are predominantly patient characteristics. While this is an appropriate approach for a tool intended to characterize surgeon and hospital performances and support operative decision making, the potential value of intraoperative information had yet to be demonstrated conclusively. Our work showed a sizable, consistent gain in the model performance by adding the intraoperative variables. With the integration of such a model with an electronic health record system and mapping of pertinent variables, it may be possible to implement a dynamic risk model that updates the predicted risk for individual patients as the data become available. Our work can serve as a prototype for future endeavors.

Second, although machine learning models have been evaluated in large clinical registries, it has not been applied to the contemporary national STS ACSD with an extensive

number of variables, and the potential advantage of machine learning approach applied to this commonly utilized dataset had been unknown. Our work demonstrated that performance gain occurred with the XGBoost approach compared with the logistic regression counterparts in 4 of the 7 evaluated outcomes. Even when the gains occurred, those measured by AUROC, Brier score, and AUPRC were small. However, when evaluating risk re-stratification across the pre-specified risk strata, the XGBoost model showed more accurate stratification. For example, for operative mortality, the logistic regression model overestimated risk in 10.6% of the patients, while overestimation was 0% in the XGBoost model. In examining several patients who had a discrepant predicted probability of mortality between XGBoost and logistic regression models, we noted that discrepancies may occur in patients with extreme observed value in variables that are key predictors of the event. This may be because algorithms had different ways of processing extreme values and that larger magnitude of predicted risks tends to have larger error margin.

The STS ACSD, despite being one of the most extensive clinical registries in cardiac surgery, constrains variables at the time of data collection, resulting in the categorization of the majority of the variables with only a small fraction retained as continuous variables. This likely limited the performance of the XGBoost algorithm, as a strength of the algorithm lies in better handling of extensive interactions between continuous variables in a high-dimensional space[8]. Therefore, the dataset likely did not allow the machine learning algorithm to realize its full potential in improving prediction, and emphasizes the importance of future work leveraging the rich health data that exist with electronic medical record systems[22].

Additionally, a continuous calibration plot suggested that the XGBoost model may improve the prediction of those at extremely high risks in all outcomes. Although the confidence interval is wide at high-risk ranges due to most events having low incidences, the

pattern appeared to be consistent that XGBoost follows the line of ideal calibration more closely than the logistic regression model counterparts did. This may have implications in better estimating the risks of those patients at extreme risks, which surgeons encounter rarely but may not have been predicted as well with existing models.

This study has several implications. First, as the predictions appear to improve with incorporating variables from multiple phases of care (preoperative and intraoperative), data acquisition, processing, and output platforms must evolve to shorten the time gap between the origination of the data and outputting the prediction in order for this improved prediction to make clinical impact[22, 23]. Given the large number of variables used, potential implementation solutions likely require integration of the prediction algorithm into the electronic medical record system. With appropriate data mapping, the model may continue to update prediction as the variable becomes available in the electronic medical record system. Second, as XGBoost may yield better prediction, not only by the conventional performance metrics but also by risk re-stratifications, especially for those with extremely high risk, widely used prediction models may benefit from adopting such a modeling approach.

Our analysis brought forth several intraoperative variables that were associated with adverse outcomes. For mortality, appropriate timing of antibiotic use and the use of antegrade cardioplegia compared to retrograde cardioplegia alone was associated with lower odds of operative mortality. As have been demonstrated, intraoperative body temperature and glucose levels were also predictive of mortality risk. As we did not evaluate such relationships in a rigorous causal inference framework, future works should determine whether they are markers or mediators of adverse outcomes. For example, the observation that transfusions are associated with increased risk of mortality may be a marker of severe conditions requiring transfusion, rather than the transfusion itself affecting short-term mortality. Similarly, the

inability to use the left internal mammary artery may represent the surgeon-specific or process-related risk and not the immediate physiologic effect towards mortality. Evaluating causal pathways for the intraoperative variables will likely improve the dynamic prediction of risk.

Finally, while not the main aim of this study, both XGBoost and logistic regression models yielded similar sets of variables that were deemed important or statistically significant for the prediction. These include preoperative hemodynamic acuity, transfusion-related variables, peak intraoperative glucose level, and intraoperative echocardiographic findings.

Limitations

Our study shares limitations of the STS ACSD risk models in that the models were developed from the dataset of those who underwent the operation. Therefore, even the models using only the preoperative variables do not encompass the entire population of potential operative candidates, only some of whom will undergo the operation. Additionally, to evaluate intraoperative predictors, we excluded intraoperative deaths. Although this was an extremely small population, this difference in the definition of operative mortality, and consequently, the slight difference in event incidences, compared with the STS ACSD risk model, should be acknowledged. Tree-based models, including XGBoost used in this study, does not yield covariate coefficients as logistic regression models do. This limits the interpretation of the relationship between covariates and outcomes in the way with which the clinical community is familiar. In order to provide the interpretability of our results, we provided lists of variables that were deemed important by the XGBoost models. Finally, the dataset was studied retrospectively and although the phase of care to which the variable set belonged to was

informed by the STS ACSD data definition, we did not have the knowledge of which variables were actually available to clinicians preoperatively and intraoperatively.

Conclusions

In predicting 7 outcomes commonly used to measure cardiac surgical outcomes, the addition of intraoperative variables to preoperative variables resulted in improved predictions of all outcomes. For most outcomes, the XGBoost model performed better than logistic regression counterparts, although the gain associated with the modeling technique was small when measured by calibration and discrimination metrics. Calibration plot and risk reclassification further demonstrated the potential advantage of the XGBoost approach. In an environment where high dimensional data can be processed, risk models based on XGBoost may provide a better prediction of adverse events to guide clinical care.

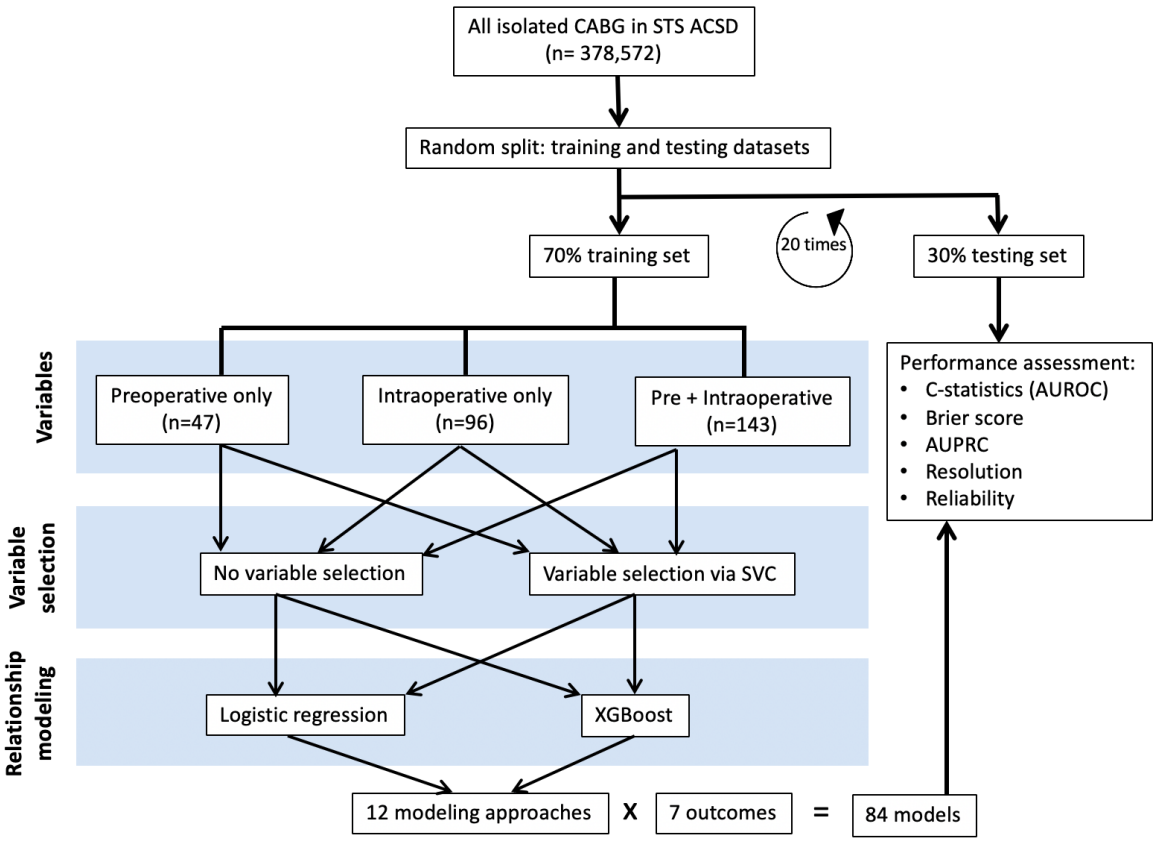
References

1. Shahian DM, O'Brien SM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1--coronary artery bypass grafting surgery. *The Annals of thoracic surgery*. Jul 2009;88(1 Suppl):S2-22. doi:10.1016/j.athoracsur.2009.05.053
2. Shahian DM, Jacobs JP, Badhwar V, et al. The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 1-Background, Design Considerations, and Model Development. *Ann Thorac Surg*. May 2018;105(5):1411-1418. doi:10.1016/j.athoracsur.2018.03.002
3. Aronson S, Stafford-Smith M, Phillips-Bute B, et al. Intraoperative systolic blood pressure variability predicts 30-day mortality in aortocoronary bypass surgery patients. *Anesthesiology*. Aug 2010;113(2):305-12. doi:10.1097/ALN.0b013e3181e07ee9
4. Celi LA, Marshall JD, Lai Y, Stone DJ. Disrupting Electronic Health Records Systems: The Next Generation. *JMIR Med Inform*. Oct 2015;3(4):e34. doi:10.2196/medinform.4192
5. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*. 2016;6
6. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *npj digital medicine*. 2018;1(18)
7. Huang C, Murugiah K, Mahajan S, et al. Enhancing the prediction of acute kidney injury risk after percutaneous coronary intervention using machine learning techniques: A retrospective cohort study. *PLoS Med*. 11 2018;15(11):e1002703. doi:10.1371/journal.pmed.1002703
8. Mortazavi BJ, Bucholz EM, Desai NR, et al. Comparison of Machine Learning Methods With National Cardiovascular Data Registry Models for Prediction of Risk of Bleeding After Percutaneous Coronary Intervention. *JAMA Netw Open*. Jul 2019;2(7):e196835. doi:10.1001/jamanetworkopen.2019.6835
9. Jacobs JP, Shahian DM, He X, et al. Penetration, Completeness, and Representativeness of The Society of Thoracic Surgeons Adult Cardiac Surgery Database. Multicenter Study. *The Annals of thoracic surgery*. Jan 2016;101(1):33-41; discussion 41. doi:10.1016/j.athoracsur.2015.08.055
10. Shahian DM, Jacobs JP, Edwards FH, et al. The society of thoracic surgeons national database. *Heart*. Oct 2013;99(20):1494-501. doi:10.1136/heartjnl-2012-303456
11. The Society of Thoracic Surgeons Adult Cardiac Surgery Database Data Collection. Web page. Jan 1 2018
12. Shahian DM, O'Brien SM, Filardo G, et al. The Society of Thoracic Surgeons 2008 Cardiac Surgery Risk Models: Part 1-Coronary Artery Bypass Grafting Surgery. *Annals of Thoracic Surgery*. Jul 2009;88(1):S2-S22. doi:10.1016/j.athoracsur.2009.05.053
13. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016; San Francisco, California, USA. 2939785
14. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. Apr 1982;143(1):29-36. doi:10.1148/radiology.143.1.7063747
15. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. presented at: Proceedings of the 23rd international conference on Machine learning; 2006; Pittsburgh, Pennsylvania, USA. 1143874
16. Brier GW. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*. 1950;78(1):1-3. doi:10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2

17. Murphy AH. A New Vector Partition of the Probability Score. *Journal of Applied Meteorology*. 1973;12(4):595-600. doi:10.1175/1520-0450(1973)012<0595:anvpot>2.0.co;2
18. Toth Z, Talagrand O, Zhu Y. The attributes of forecast systems: a general framework for the evaluation and calibration of weather forecasts. *Predictability of Weather and Climate* 584 (2006): 595.
19. Finazzi S, Poole D, Luciani D, Cogo PE, Bertolini G. Calibration belt for quality-of-care assessment based on dichotomous outcomes. *PLoS One*. Feb 2011;6(2):e16110. doi:10.1371/journal.pone.0016110
20. Pepe MS, Fan J, Feng Z, Gerds T, Hilden J. The Net Reclassification Index (NRI): a Misleading Measure of Prediction Improvement Even with Independent Test Data Sets. *Stat Biosci*. Oct 2015;7(2):282-295. doi:10.1007/s12561-014-9118-0
21. Pedregosa F, Ga, #235, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
22. Mori M, Schulz WL, Geirsson A, Krumholz HM. Tapping Into Underutilized Healthcare Data in Clinical Research. *Ann Surg*. Aug 2019;270(2):227-229. doi:10.1097/SLA.0000000000003329
23. Vahi K, Harvey I, Samak T, et al. A Case Study into Using Common Real-Time Workflow Monitoring Infrastructure for Scientific Workflows. journal article. *Journal of Grid Computing*. September 01 2013;11(3):381-406. doi:10.1007/s10723-013-9265-4

Tables and Figures

Figure 1: Analysis flow for development and evaluation of models



CABG = coronary artery bypasss graft surgery; STS ACSD= Society of Thoracic Surgeons Adult Cardiac Surgery Database; AUPRC = area under the precision-recall curve

The figure summarizes the modeling approach and metrics used to evaluate the performance. Combinations of variable sets, variable selection approach, and modeling technique for 7 outcomes resulted in 84 different models.

Figure 2: Model performances for mortality and renal failure

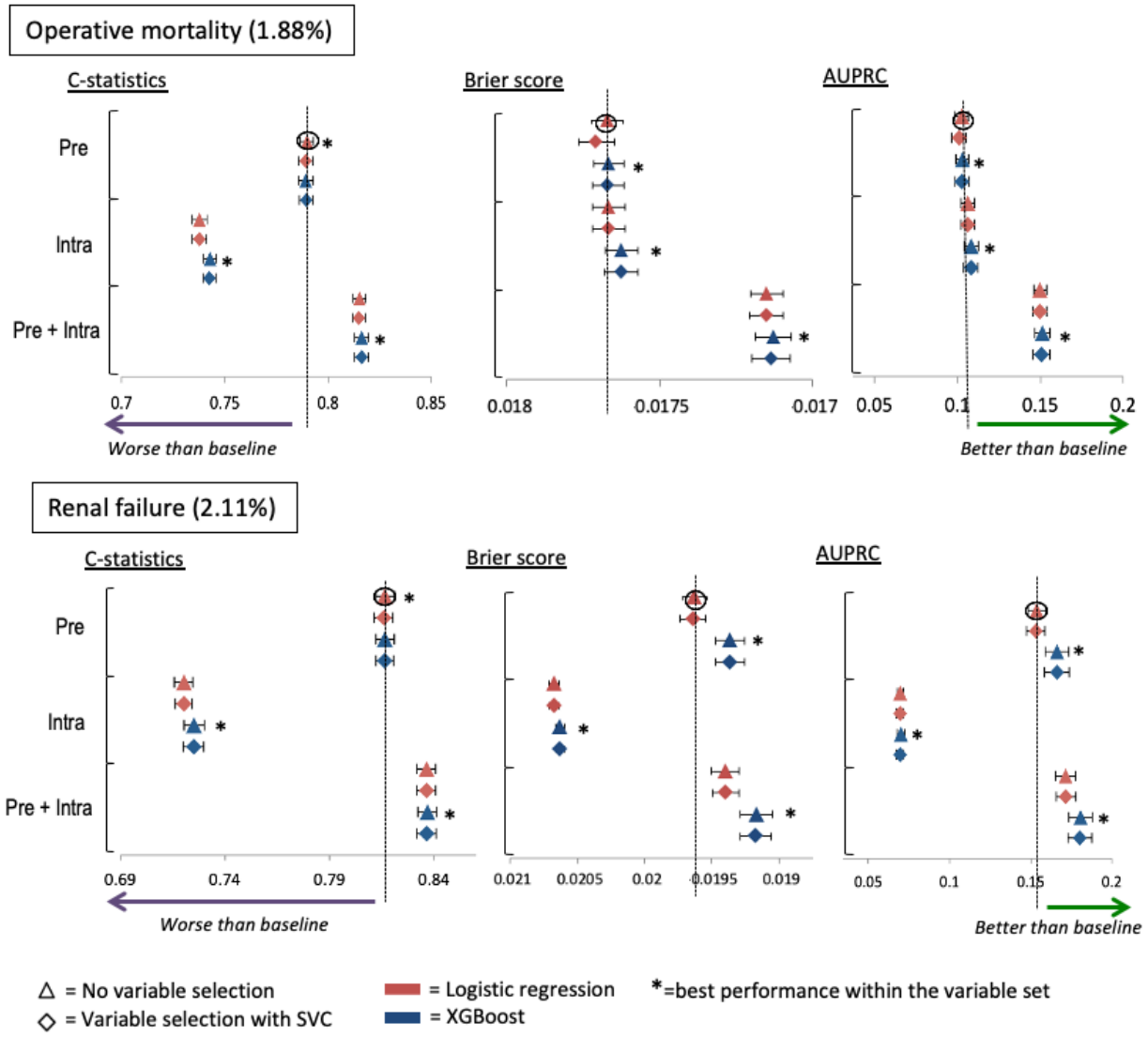


Figure summarizes model performances for mortality and renal failure evaluated by 3 metrics. Circled red triangles represent the baseline model, which are logistic regression models using preoperative variables only without further variable selection. * indicates the model with best performance within the same variable set. For all metrics, the right end of the x-axis is better and the left end is worse. For example, for operative mortality, XGBoost model using pre+intraoperative variables without variable selection had the best performance in c-statistics, Brier score, and the area under precision-recall curve (AUPRC).

Figure 3: Continuous calibration plot for 7 outcomes

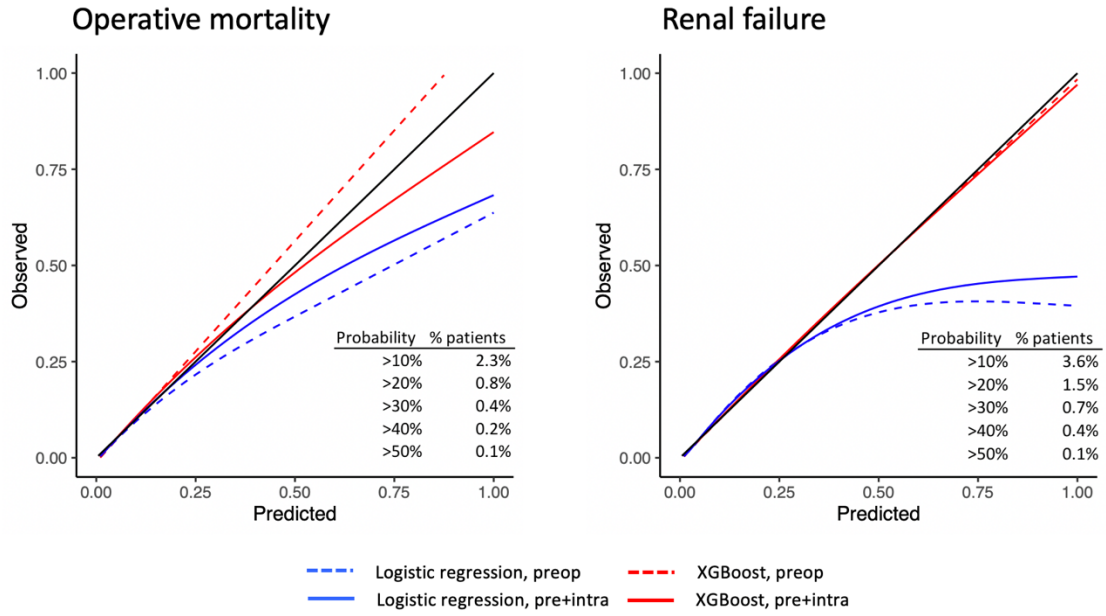


Figure shows continuous calibration plot for mortality (left) and renal failure (right). Red lines are XGBoost and blue lines are logistic regression model calibrations. Dotted lines are models using preoperative variables only and solid lines are models using pre+intraoperative variables. Black line represents perfect calibration. The legend shows percent of the cohort that had predicted event probability above the indicate threshold in percentage. For example, for operative mortality, 2.3% of the patients had predicted probability of operative mortality >10%.

Figure 4: Shift table of predicted risk for operative mortality: Logistic regression with preoperative variables vs. XGBoost with pre+intraoperative variables

	Base Model (predicted risk)					
Best Model (predicted risk)	**<1%[1]** Event rate (No. patients)	**1-3%[2]** Event rate (No. patients)	**3-5%[3]** Event rate (No. patients)	**5-10%[4]** Event rate (No. patients)	**>10%[5]** Event rate (No. patients)	**All** Event rate (No. patients)
<1%[a]	0.4% (51124)	0.9% (10995)	0% (34)	0% (4)	0% (0)	0.5% (62157)
1-3%[b]	1.1% (5582)	1.7% (26157)	3.1% (4336)	5.1% (634)	16.7% (12)	1.8% (36721)
3-5%[c]	5.9% (171)	3.7% (2454)	3.5% (2804)	5.1% (1938)	7.9% (127)	4.1% (7494)
5-10%[d]	9.6% (73)	8.2% (644)	8.1% (973)	6.9% (2032)	8.2% (845)	7.6% (4567)
>10%[e]	25.0% (48)	22.5% (271)	19.2% (297)	19.0% (601)	20.1% (1416)	20.1% (2633)
All	0.5% (56998)	1.8% (40521)	4.4% (8444)	7.4% (5209)	15.2% (2400)	1.9% (113572)

The figure shows predicted risk of operative mortality by the base model (logistic regression using preoperative variables without variable selection) and the best model (XGBoost using pre+intraoperative variables without variable selection). Actual observed mortality rate is indicated by the % and numbers in parenthesis indicate number of all patients in each predicted risk strata. Gray cells are those classified in the same stratum by both models. Base model underestimated 11,114 patients (9.8%) and overestimated 12,005 patients (10.6%). Best model underestimated 7,218 patients (6.4%) and overestimated 0 patients (0%).
*1-5 denotes cell location by the column and a-e denotes cell location by the row.

Figure 5: Shift table of predicted risk for operative mortality: Logistic regression with pre+intraoperative variables vs. XGBoost with pre+intraoperative variables

	Base Model (predicted risk)					
	<1%[1]	**1-3%[2]**	**3-5%[3]**	**5-10%[4]**	**>10%[5]**	**All**
Best Model (predicted risk)	Event rate (No. patients)	Event rate (No. patients)	Event rate (No. patients)	Event rate (No. patients)	Event rate (No. patients)	Event rate (No. patients)
<1%[a]	0.5%(58696)	0.9% (3454)	0% (6)	0% (1)	0% (0)	0.5% (62157)
1-3%[b]	1.1% (5867)	1.8% (28650)	3.1% (2093)	4.8% (105)	16.7% (6)	1.8% (36721)
3-5%[c]	0% (5)	2.8% (1881)	4.0% (4289)	6.3% (1296)	13.0% (23)	4.1% (7494)
5-10%[d]	0% (0)	13.1% (61)	6.1% (847)	7.2% (2980)	10.9% (679)	7.6% (4567)
>10%[e]	0% (0)	0% (0)	11.1% (9)	12.2% (353)	21.3% (2271)	20.1% (2633)
All	0% (0)	1.8% (34046)	4.0% (7244)	7.3% (4735)	18.9% (2979)	1.9% (113572)

The figure shows predicted risk of operative mortality by the base model (logistic regression using preoperative variables without variable selection) and the best model (XGBoost using pre+intraoperative variables without variable selection). Actual observed mortality rate is indicated by the % and numbers in parenthesis indicate number of all patients in each predicted risk strata. Gray cells are those classified in the same stratum by both models. Base model underestimated 7,137 patients (6.3%) and overestimated 3,566 patients (3.1%). Best model underestimated 4,263 patients (3.8%) and overestimated 1,886 patients (1.7%).
*1-5 denotes cell location by the column and a-e denotes cell location by the row.

Tables:

Table 1: Predictor variables

Preoperative variables	Type or level
Age	Linear spline with knots at 50 and 60 years, interaction term with case status and incidence of operation
Race/ethnicity	Caucasian, Black, Asian, Hispanic
Sex	Female, Male
Body surface area	Continuous, conditioned on sex
Chronic lung disease	Mild, Moderate, Severe
Last preoperative creatinine level	Linear spline with knots at 1.0 and 1.5, conditioned on preoperative dialysis
Preoperative dialysis	Yes, No
Hypertension	Yes, No
Diabetes	No, Non-insulin dependent, insulin dependent
Congestive heart failure	Yes, No, conditioned on NYHA class
Cerebrovascular disease	Yes, No, conditioned on stroke/TIA
Peripheral vascular disease	Yes, No
Atrial fibrillation	Yes, No
Immunosuppressed status	Yes, No
Left main disease	Yes (≥ 50% stenosis), No
Myocardial infarction	No, within 6 hours, 24 hours, 21 days, ≥21 days
Number of diseased vessels	Discrete continuous
PCI < 6 hours	Yes, No
Shock	Yes, No
Inotrope or IABP use preoperatively	Yes, No
Prior cardiovascular surgery	None, once, ≥ twice
Ejection fraction	Continuous
Mitral insufficiency	Yes, No (Yes = moderate or severe)
Tricuspid insufficiency	Yes, No (Yes = moderate or severe)
Aortic stenosis	Yes, No (Yes = moderate or severe)
Case status	Elective, Urgent, Emergent, Salvage

Intraoperative variables	Type or level
Operative approach	Sternotomy, thoracotomy, partial sternotomy, port access
Conversion of planned approach	Yes, No
Robot used	Yes, No
Prophylactic antibiotics used	Yes, No
Antibiotics given within 1 hour of incision	Yes, No
Prophylactic antibiotics redosed	Yes, No
Lowest body temperature	Continuous
Lowest hemoglobin, hematocrit	Continuous
Cardiopulmonary bypass use	None, Combination, Full, reasons if combination
Amino caprioc acid use	Yes, No
Tranexamic acid use	Yes, No
Clotting factor use	Yes, No

Cardioplegia delivery	Antegrade, Retrograde, Both
Cardioplegia type	Blood, Crystalloid, Both, Other
Blood product use	Yes, No
Unit of red blood cell transfused	Continuous
Unit of platelet transfused	Continuous
Unit of fresh frozen plasma transfused	Continuous
Unit of cryoprecipitate transfused	Continuous
Intraoperative transesophageal echo	Ejection fraction, valvular insufficiency
Time from skin incision to closure	Continuous

The table summarizes predictor variables by general categories.

Table 2: Incidences of Postoperative Events

Events	**N=378,572**
Operative mortality	1.88%
Permanent stroke	1.31%
Renal failure	2.11%
Prolonged ventilation	8.05%
Reoperation	3.47%
DSWI	0.30%
Composite morbidity and mortality	12.08%

*DSWI = deep sternal wound infection

Table 3: Odds ratio estimate for variables selected by support vector machine in predicting mortality

Variables	OR	2.50%	97.50%
Preoperative			
Body surface area	2.464	2.011	3.019
Shock	2.396	2.144	2.677
Dialysis	2.066	1.67	2.556
Severe lung disease	2.013	1.852	2.188
Salvage status	2.004	1.712	2.345
IABP or inotrope dependent	1.661	1.539	1.791
Creatinine (splined at 1.0)	1.512	1.194	1.914
CHF with NYHA III/IV	1.477	1.36	1.604
MI (within 24 hours)	1.469	1.303	1.657
Tricuspid insufficiency	1.463	1.335	1.604
Peripheral vascular disease	1.439	1.356	1.526
Moderate lung disease	1.331	1.207	1.467
CHF with NYHA I/II	1.324	1.245	1.408
Immunosuppressed	1.309	1.173	1.461
CVD without stroke	1.293	1.198	1.397
MI (within 21 days)	1.232	1.163	1.304
Mitral insufficiency	1.206	1.121	1.299
Urgent status	1.145	1.076	1.219
Insulin-dependent diabetes	1.142	1.074	1.214
Number of diseased vessels	1.135	1.074	1.199
CVD with stroke	1.127	1.054	1.206
Age (splined at 50)	1.116	1.095	1.138
Left main disease	1.079	1.025	1.136
Age*Case urgency	1.023	1.019	1.028
Age*Reoperative status	1.011	1.006	1.017
Age (splined at 60)	1.003	0.989	1.017
Ejection fraction	0.982	0.978	0.987
Age	0.933	0.924	0.943
[#]Creatinine	0.896	0.722	1.111
Intraoperative			
Intraoperative IABP	4.419	4.054	4.816
Any blood products used	1.168	1.095	1.247
pRBC (number of units used)	1.122	1.103	1.142
Intraoperative TEE performed	1.093	1.027	1.163
FFP (number of units used)	1.066	1.04	1.092
Highest intraoperative glucose level	1.002	1.002	1.002
Skin-to-skin time	1.002	1.002	1.002
Lowest body temperature	0.98	0.97	0.989

Blood cardioplegia use	0.927	0.868	0.99
Lowest intraoperative hemoglobin level	0.924	0.909	0.94
Number of distal anastomosis	0.902	0.878	0.928
Postop EF (increased)	0.88	0.82	0.945
LIMA used	0.844	0.781	0.911
#Appropriate antibiotics used	0.844	0.696	1.022
Antegrade cardioplegia	0.822	0.764	0.884
Postop tricuspid regurgitation (None)	0.812	0.75	0.879
Appropriate timing of antibiotics use	0.793	0.65	0.967
Antegrade and retrograde cardioplegia	0.783	0.727	0.843
Postop tricuspid regurgitation (trace/trivial)	0.782	0.718	0.851
Full sternotomy	0.779	0.64	0.949
Planned use of combination CPB	0.585	0.474	0.721

#Indicates variables with confidence interval crossing 1.0.

Table 4: Variables on the order of importance in XGBoost model for mortality

Intraoperative variables	Importance score
Plasma transfusion	291.7
Any transfusion	219.9
Intraoperative IABP use	181.9
Red cell transfusion	122.8
Internal thoracic artery use	50.0
Lowest hemoglobin level	40.2
Postop ejection fraction	39.3
Highest glucose level	34.7
Surgery duration	30.7
Cardioplegia delivery approach	27.2
Postop tricuspid insufficiency	15.4
Number of distal anastomosis	15.0
Postop mitral insufficiency	14.1
Unplanned use of combination CPB	13.7
Platelet transfusion	13.6
Tranexamic acid use	11.3
Postop aortic insufficiency	10.1
Operative approach (full/partial sternotomy, thoracotomy)	9.1
Cryoprecipitate transfusion	8.5
Lowest body temperature	8.3
Whether additional prophylactic antibiotic dose given	7.9
Timing of antibiotics dosing	7.3
Clotting factor administration	5.9
Preoperative variables	
Preop IABP or inotrope use	193.2
Shock	166.0
CHF*NYHA class	165.9
Peripheral vascular disease	115.9
Age*Status	97.8
Chronic lung disease	92.3
Tricuspid insufficiency	89.7
Mitral insufficiency	85.8
Ejection fraction	85.2
Age	63.6
Creatinine	59.7
Timing of myocardial infarction	43.0
Status (elective, urgent, emergent, salvage)	40.2
Insulin-dependent diabetes	19.5
Sex*body surface area	18.9
Age*Redo sternotomy	17.3
Number of diseased vessels	16.8

Body surface area	14.1
Stroke	13.8
PCI within 6 hours	12.0
Left main disease	11.6
Immunosuppressed status	10.4
Hypertension	10.4
Atrial fibrillation	9.9
Aortic stenosis	9.9
Transient ischemic attack	8.6
Race	7.5
Ethnicity	6.0

CPB=cardiopulmonary bypass; IABP = intraaortic balloon pump; CHF = congestive heart failure; NYHA = New York Heart Association; PCI = percutaneous coronary intervention. *indicates interaction between the two variables. Importance score denotes the average number of times the variable was used to split the trees in XGBoost.

Supplementary Table S1: Final XGBoost model parameters

Events	Number of estimators	Maximum tree depth	Learning rate	C
Operative mortality	257	2	0.3	0.5
Permanent stroke	182	2	0.4	0.15
Renal failure	220	3	0.3	0.125
Prolonged ventilation	380	3	0.3	0.2
Reoperation	178	3	0.4	0.175
DSWI	92	2	0.4	0.8
Composite morbidity and mortality	324	3	0.3	0.5

DSWI = deep sternal wound infection

Table summarizes two parameters that were tuned for XGBoost models. C values are only applicable for models with variable selection, and represent penalty value to regularize variable selection in the support vector machine algorithm, with 0 exerting maximum regularization and 1 exerting no regularization.

Supplementary Table S2: Shift table for renal failure using pre+intraoperative variables comparing XGBoost and logistic regression

	Base Model					
Best Model	<1% Event rate (No. patients)	1-3% Event rate (No. patients)	3-5% Event rate (No. patients)	5-10% Event rate (No. patients)	>10% Event rate (No. patients)	All Event rate (No. patients)
<1%	(0.5%) 56455	(0.8%) 5886	(0%) 4	(0%) 0	(0%) 0	(0.48%) 62345
1-3%	(1.0%) 2926	(1.8%) 26665	(2.9%) 2430	(6.1%) 131	(50.0%) 2	(1.8%) 32154
3-5%	(0%) 14	(2.5%) 1513	(3.6%) 3790	(6.5%) 1332	(12.8%) 39	(4.0%) 6688
5-10%	(33.3%) 3	(3.3%) 123	(5.1%) 1103	(8.0%) 3107	(12.1%) 595	(7.8%) 4931
>10%	(0%) 1	(0%) 5	(11.4%) 44	(11.8%) 845	(22.4%) 3097	(20.0%) 3992
All	(0.5%) 59399	(1.7%) 34192	(3.7%) 7371	(8.2%) 5415	(20.7%) 3733	(2.12%) 110110

The table shows predicted risk of renal failure by the base model (logistic regression using pre+intraoperative variables without variable selection) and the best model (XGBoost using pre+intraoperative variables without variable selection). Actual observed renal failure rate is indicated by the % and numbers in parenthesis indicate number of all patients in each predicted risk strata. Gray cells are those classified in the same stratum by both models. Base model underestimated 5,044 patients (4.6%) and overestimated 8,325 patients (7.6%). Best model underestimated 2,102 patients (1.9%) and overestimated 1,656 patients (1.5%).

Supplementary Table S3: Intraoperative variables

Intraoperative variables	N, median	%, IQR
Total	378572	
Appropriate antibiotics selected		
Yes	373358	98.6%
No	1498	0.4%
Exclusion (reason for different selection known)	3716	1.0%
Appropriate timing of antibiotics administration		
Yes	374124	98.8%
No	2431	0.6%
Exclusion (contraindication)	2017	0.5%
Additional prophylactic antibiotics dose given	179023	47.3%
Lowest temperature	34	33-35.1
Lowest hemoglobin	8.9	7.8-9.5
Operative approach converted during surgery	5770	1.5%
Missing	103	0.0%
CPB Use		
None	50889	13.4%
Combination	4212	1.1%
Full	327683	86.6%
Missing	407	0.1%
Reasons for combined CPB utilization (off-pump to on-pump)		
Exposure	281	0.1%
Bleeding	39	0.0%
Inaddquate size/diffuse disease	95	0.0%
Hemoduynamic instability	1123	0.3%
Conduit quality/trauma	50	0.0%
Other	126	0.0%
Missing/No combination CPB	376858	99.5%
Cardioplegia delivery		
None	64809	17.1%
Antegrade	167376	44.2%
Retrograde	3229	0.9%
Both	140615	37.1%
Missing	2543	0.7%

Cardioplegia type		
Blood	247317	65.3%
Crystalloid	17391	4.6%
Both	46227	12.2%
Other/missing	2828	0.7%
None	64809	17.1%
Intraop use of IABP	6980	1.8%
IMA use		
Left	340535	90.0%
Right	2775	0.7%
Both	18383	4.9%
No IMA	16875	4.5%
Missing	4	0.0%
Epsilon Amino-Caproic Acid use (Yes)	274896	72.6%
Missing	142	0.0%
Tranexamic acid use	37232	9.8%
Missing	140	0.0%
Intraop clotting factor use	4987	4987
Intraop TEE	208365	55.0%
Number of distal anastomosis	3	3-4
Operative approach		
Full sternotomy	372767	98.5%
Partial sternotomy	1311	0.3%
Right or left parasternal incision	97	0.0%
Left thoracotomy	862	0.2%
Right thoracotomy	30	0.0%
Transverse sternotomy	25	0.0%
Sub-xiphoid	26	0.0%
Sub-costal	12	0.0%
Bilateral thoracotomy	28	0.0%
Mini-thoracotomy	2819	0.7%
Port-access	308	0.1%
Missing	287	0.1%
Postop aortic insufficiency		
None	109561	28.9%
Trace/trivial	25386	6.7%
Mild	12209	3.2%
Moderate	2304	0.6%

Severe	115	0.0%
Not reported	58790	15.5%
No TEE obtained/missing	170207	45.0%
Postop ejection fraction		
Unchanged	97823	25.8%
Increased	63570	16.8%
Decreased	7089	1.9%
Not reported	39883	10.5%
No TEE obtained/missing	170207	45.0%
Postop mitral insufficiency		
None	47130	12.4%
Trace/trivial	59137	15.6%
Mild	43034	11.4%
Moderate	43034	11.4%
Severe	523	0.1%
Not reported	50012	13.2%
No TEE obtained/missing	135702	35.8%
Postop tricuspid insufficiency		
None	65832	17.4%
Trace/trivial	53122	14.0%
Mild	23334	6.2%
Moderate	3704	1.0%
Severe	422	0.1%
Not reported	61953	16.4%
No TEE obtained/missing	170205	45.0%
Blood product use		
Yes	103794	27.4%
No	272703	72.0%
Refused	2075	0.5%
Missing	0	0.0%
Transfusion units		
Cryoprecipitate	0	0-0
Fresh frozen plasma	0	0-0
Platelet	0	0-0
Red blood cell	0	0-0
Robot use	3364	0.9%
Skin incision duration	225	182-275

CPB= cardiopulmonary bypass; IMA= internal mammary artery; IABP=intraaortic balloon pump; TEE= transesophageal echocardiogram.

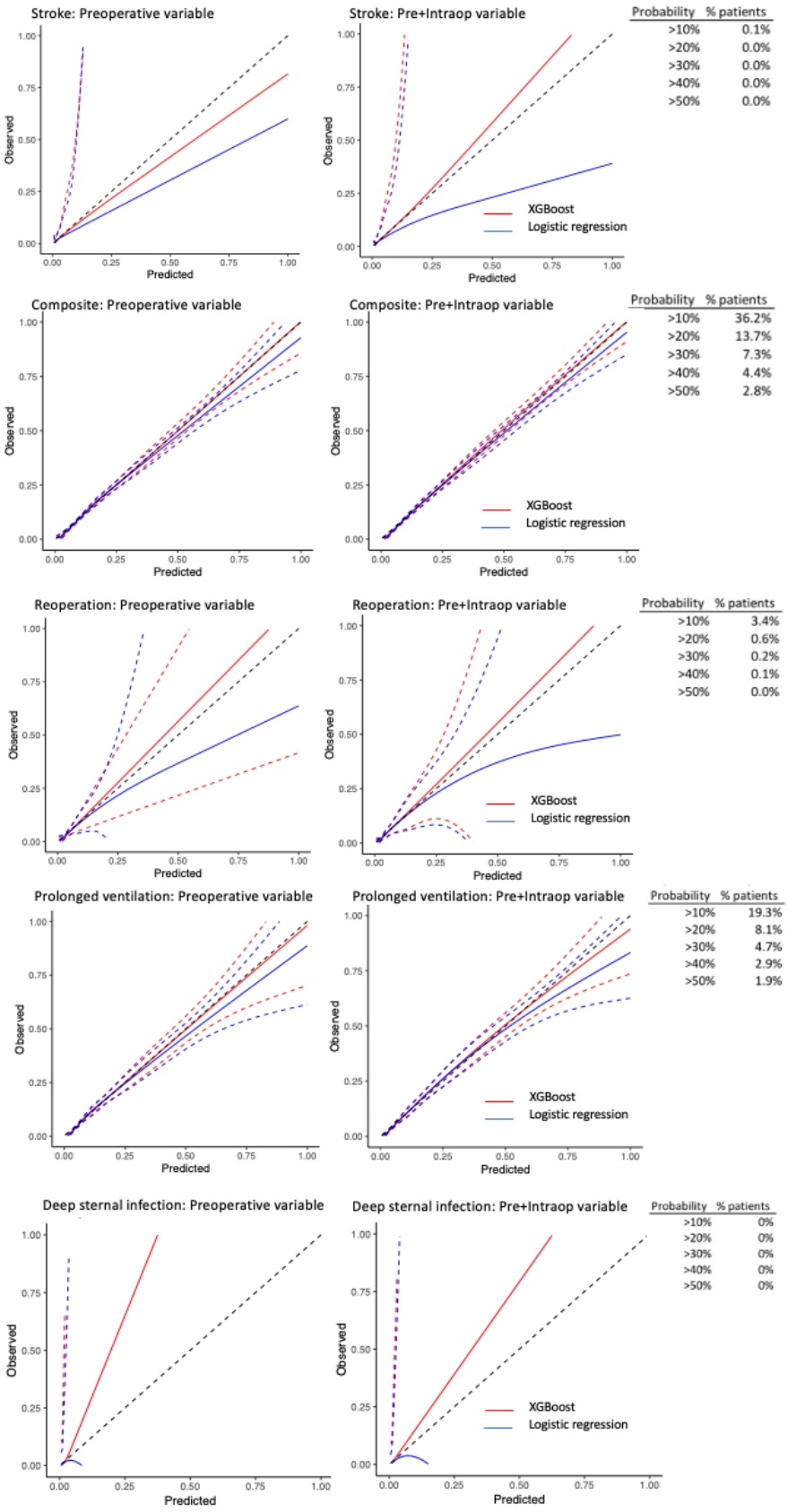
Supplementary Table S5: Combinations of variables, modeling technique, and variable selection

Variable sets	Relationship model	Variable selection	Label
Preoperative	Logistic regression	None	*1
		SVC	2
	XGBoost	None	3
		SVC	4
Intraoperative	Logistic regression	None	5
		SVC	6
	XGBoost	None	7
		SVC	8
Pre + Intraoperative	Logistic regression	None	9
		SVC	10
	XGBoost	None	11
		SVC	12

SVC = support vector classifier

*Model 1 is considered the baseline model for comparisons.

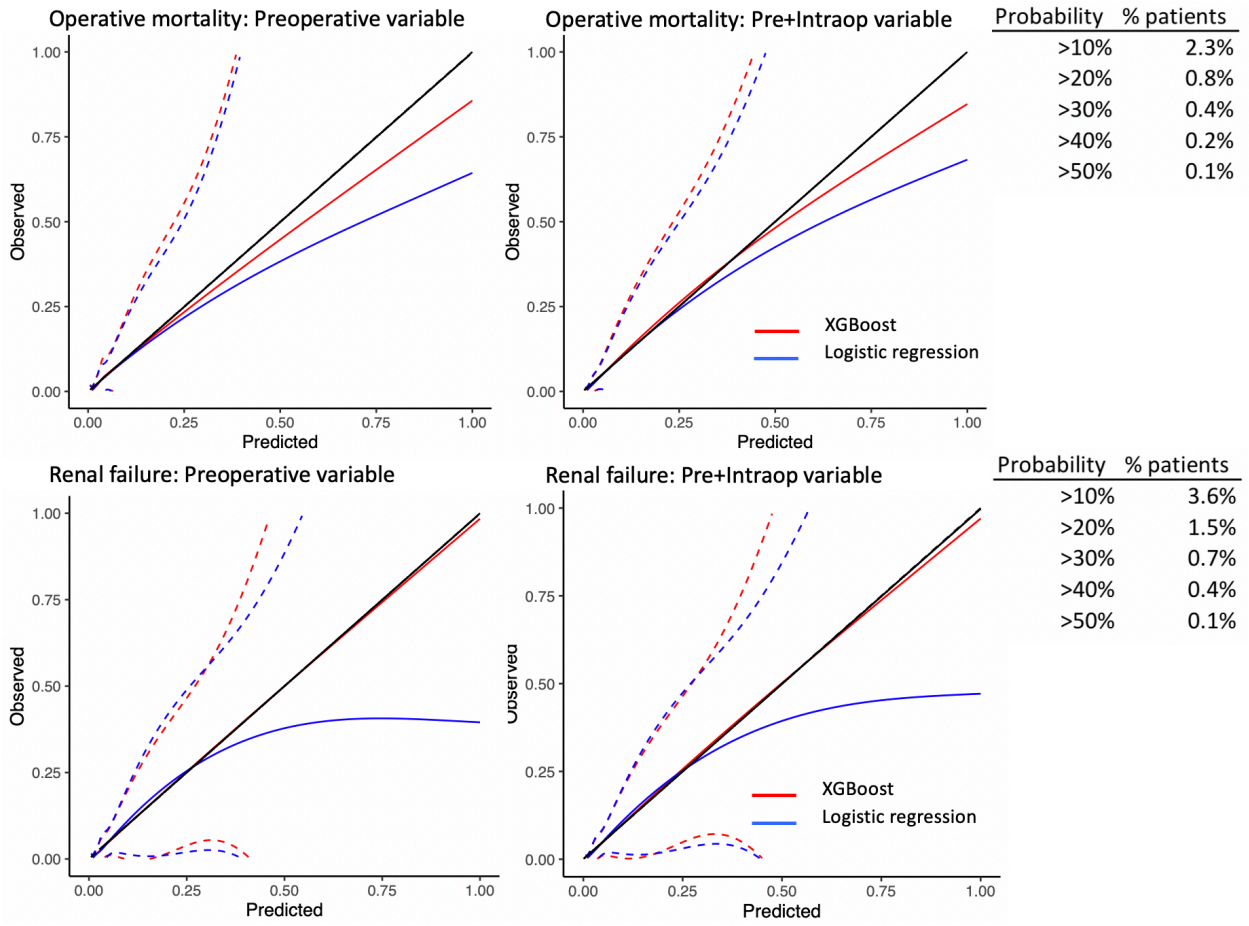
Supplementary Figure S1: Continuous calibration plot



Figures shows continuous calibration plot for the indicated outcomes. Red solid line is XGBoost and blue solid line is logistic regression model calibrations. Dotted colored lines are 95%

confidence interval for corresponding models. Dotted black line represents perfect calibration. Right legend shows percent of the cohort that had predicted event probability above the indicate threshold in percentage.

Supplementary Figure S2: Continuous calibration plot for mortality and renal failure



Figures shows continuous calibration plot for the indicated outcomes. Red solid line is XGBoost and blue solid line is logistic regression model calibrations. Dotted colored lines are 95% confidence interval for corresponding models. Dotted black line represents perfect calibration. Right legend shows percent of the cohort that had predicted event probability above the indicate threshold in percentage.

Supplementary Text: Case examples of predicted risks of death that were divergent between XGboost and logistic regression models

In examining several patients who had discrepant predicted probability of mortality between XGBoost and logistic regression models, we noted that discrepancies may occur in patients with extreme observed value in variables that are key predictor of the event. For example, a 42-year-old man with minimal comorbidity who underwent emergent 3-vessel CABG received 23 units of red blood cell and 16 units of plasma. This patient had 44% and 6% chance of death predicted by XGBoost and logistic regression, respectively. Similarly, a 70-year-old man who underwent salvage 5-vessel CABG with intraoperative IABP placement. XGBoost and logistic regression models predicted the risk of mortality of this patient to be 30% and 10%, respectively. In a case where logistic regression had a significantly higher predicted mortality probability, a 52-year-old woman with minimal comorbidity undergoing elective 5-vessel CABG required IABP during the operation and had the lowest body temperature of 27°C. This patient had predicted mortality of 6% by XGBoost and 20% by logistic regression.

CHAPTER 5

Concluding Remarks

Understanding the patterns of postoperative recovery after cardiac surgery is important from several perspectives: to facilitate patient-centered treatment decision making, to inform health care policy targeted to improve postoperative recovery, and to guide patient care after cardiac surgery. Although existing literature has described postoperative recovery after cardiac surgery, we demonstrated through our systematic review that the current approaches to measuring and reporting recovery as a treatment outcome varies widely across studies. This made synthesis of collective knowledge challenging. Notably, there were no standards regarding the number of measurements taken over variable follow-up duration even for studies using the same patient-reported outcome measure instruments such as SF-36. Therefore, our systematic review highlighted key gaps in knowledge, which we sought to address in our prospective cohort study,

We conducted a prospective single-center cohort study of patients after cardiac surgery to measure their recovery trajectory across multiple domains of recovery. This study leveraged digital platform to facilitate frequent data collection over 30 days after surgery to visualize a granular evolution of patient recovery after cardiac surgery. We used a latent class analysis to facilitate identification of dominant trajectory patterns that had been obscured in a conventional way of reporting such time-series data using group-level means. For the pain domain, we identified 4 trajectory classes, one of which was a group of patients with persistently high pain trajectory that only became distinguishable from less concerning group after 10 days. This information is potentially useful in tailoring individualized follow-up timing after surgery to improve the pain control.

The prospective study embodied several important features to successfully conducting such studies of patient-reported outcome. This included the use of digital platform to facilitate efficient data collection even after hospital discharge, iteratively improving the protocol to optimize patient engagement including evaluation of potential barriers to survey completion, and using latent class analysis to identify dominant patterns of recovery trajectories. We outlined these insights in the protocol manuscript to inform subsequent studies aiming to leverage such a digital platform to measure longitudinal patient-centered outcome.

Finally, we evaluated the potential value of incorporating health care data generated in the different phases of patient care in improving the prediction of postoperative outcomes after cardiac surgery. The current convention is the Society of Thoracic Surgeons' risk model, which only uses patient data available preoperatively. We demonstrated that the addition of intraoperative variables to the conventional preoperative variable set improves the performance of prediction models substantially. Using machine learning approach to such a high-dimensional dataset proved to be only marginally important, however. This work demonstrated the potential value and importance of being able to leverage health care data to continuously update the prediction to inform patient outcomes and guide clinical care.

Our work collectively advanced knowledge in several key aspects of postoperative recovery. First, we highlighted the knowledge gap in the existing literature through characterizing the variability in the ways such studies had been conducted. Second, we designed and described a framework to measure postoperative recovery and an analytical approach to informatively characterize longitudinal patient recovery. Third, we employed these designs in a prospective cohort study to measure and analyze recovery trajectories and described clinical insights obtained from the study. Finally, we demonstrated the potential

value of a dynamic risk model to improve on its predictive performance by incorporating new data generated as the patient progresses through the phase of care. Such a platform has the potential to individualized approach to improve postoperative recovery.

Appendix: Bibliography of the thesis work

1. Mori M, Angraal S, Chaudhry SI, et al. Characterizing Patient-Centered Postoperative Recovery After Adult Cardiac Surgery: A Systematic Review. *J Am Heart Assoc*. 2019;8(21):e013546.

2. Mori M, Brooks C, Spatz E, et al. Protocol for project recovery after cardiac surgery: a single-center cohort study leveraging digital platform to characterise longitudinal patient-reported postoperative recovery patterns. *BMJ Open*. 2020;10(9):e036959.

3. Mori M, Bin Mahmood SU, Zhuo H, et al. Persistence of risk of death after hospital discharge to locations other than home after cardiac surgery. *J Thorac Cardiovasc Surg*. Published online March 6, 2019.

4. Mori M, Krumholz HM, Allore HG. Using Latent Class Analysis to Identify Hidden Clinical Phenotypes. *JAMA*. 2020;324(7):700-701.

5. Mori M, Khera R, Lin Z, Ross JS, Schulz W, Krumholz HM. The Promise of Big Data and Digital Solutions in Building a Cardiovascular Learning System: Opportunities and Barriers. *Methodist Debakey Cardiovasc J*. 2020;16(3):212-219.

6. Mori M, Shahian DM, Suter LG, Geirsson A, Lin Z, Krumholz HM. Relevance of Cardiac Surgery Outcome Reporting 3 Years Later in a New York and California Statewide Analysis. *JAMA Surg*. 2020;155(5):442-444.

7. Mori M, Djulbegovic M, Hajduk AM, Holland ML, Krumholz HM, Chaudhry SI. Changes in Functional Status and Health-Related Quality of Life in Older Adults After Surgical, Interventional, or Medical Management of Acute Myocardial Infarction. *Semin Thorac Cardiovasc Surg*. 2021;33(1):72-81.

8. Mori M, Nasir K, Bao H, et al. Administrative Claims Measure for Profiling Hospital Performance Based on 90-Day All-Cause Mortality Following Coronary Artery Bypass Graft Surgery. *Circ Cardiovasc Qual Outcomes*. 2021;14(2):e006644.

9. Mori M, Brooks C, Dhruva SS, et al. Trajectories of pain after cardiac surgery: implications for measurement, reporting, and individualized treatment. *Circ Cardiovasc Qual Outcomes*. 2021 [In press].

10. Mori M, Nasir K, Bao H, et al. Towards Dynamic Risk Prediction of Outcomes After CABG: Improving Risk Prediction with Intraoperative Events Using Gradient Boosting. *Circ Cardiovasc Qual Outcomes*. 2021 [In press].

11. Mori M, Schulz WL, Geirsson A, Krumholz HM. Tapping Into Underutilized Healthcare Data in Clinical Research. *Ann Surg*. 2019;270(2):227-229.