Urinary Continence Recovery After Radical Prostatectomy Using Patient Reported Outcomes Data: Variability, Predictions, and Prediction Accuracy

by

Roshan Paudel

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Health Infrastructures and Learning Systems)
in the University of Michigan
2022

Doctoral Committee:

Associate Professor Gretchen A. Piatt, Chair
Associate Professor Matthew A. Davis
Professor Brent K. Hollenbeck
Assistant Professor Andrew E. Krumm
Assistant Professor Karandeep Singh

Roshan Paudel
rpaudel@umich.edu
ORCID iD: 0000-0002-8584-7045

Dedication

This dissertation is dedicated to Misti and Michael.

Your love and support mean so much to me.

Acknowledgements

I would like to express my deepest gratitude to my dissertation committee. This dissertation would not have been possible without the guidance, support, and encouragement from Dr. Gretchen Piatt. I am forever grateful for her steady mentorship and your patience throughout the past four years. I want to extend my sincere appreciation to Dr. Karandeep Singh for his support, guidance, and for teaching me everything about R and machine learning. I have learned so much from you over the past four years. Dr. Andrew Krumm, you are truly a delight to work with. You taught me mixed-effects models and you patiently supported me through the process. You encouraged me to think about what is practical, realistic, and how my work might help someone diagnosed with prostate cancer. Dr. Brent Hollenbeck, thank you for your invaluable clinical knowledge and expertise. You pushed me to think about what my dissertation adds to the general knowledge in this field. You challenged me to think about the clinical implications and asked the tough questions. Dr. Matt Davis, you taught me epidemiology and applied biostatistics and supported me through this journey. Thank you to my committee. You have made me a better student of science.

I grateful to Dr. James Montie for welcoming me into MUSIC. To Drs. Khurshid Ghani, Kevin Ginsburg, Giulia Lane, Arvin George, Alice Semerjian, and Brian Lane – you supported me, gave me opportunities to co-author manuscripts with you and encouraged me to continue pursuing my interests in prostate cancer research. I am also thankful to Susan Linsell, Anna Johnson, Stephanie Ferrante, Ji Qi, Rod Dunn, and all others at MUSIC for their support. You all welcomed me with open arms and gave me an opportunity to learn from you. Kevin Ginsburg, Stephanie Ferrante, and Ji Qi— it has been a great honor collaborating with you on many prostate projects.

Dr. Emily Kobernik – you always supported me, encouraged me and you were always available when I needed you. Drs. Gracie Trinidad, Nikolas Koscielniak (I spelled your name correctly, Nik), Rama Mwenesi, and Elliott Brannon—thank you for leading the way. You were there when I needed your help. I am grateful to my peers including Stephanie Hall, Victor Rentes, and Anthony Provenzano for their support and camaraderie. I enjoyed being in the

program with you all. To Dr. Chuck Friedman - thank you for supporting me and for giving me career advice. To Beth Hill and Elizabeth Rodriguez, this would not have been possible without your unwavering support. Thank you!

Table of Contents

List of Tables

List of Figures

Abstract

Despite advances in prostate cancer treatment, wide variability in post-operative oncologic and functional outcomes are observed. Effective management of prostate cancer requires accurate prediction of potential outcomes and setting realistic expectations of functional outcomes is an important challenge. Advances in computing are accelerating the development of prediction tools using routine clinical practice and patient reported outcomes data to support the treatment decision-making process; however, variability in post-operative functional outcomes remains despite the advancements.

The development, deployment, and evaluation of prediction models in routine urology practices has the potential to enhance the treatment decision-making process and improve the overall quality of prostate cancer care. The objective of this dissertation is to assess the patterns of urinary continence recovery in a statewide registry. To achieve the objective of this dissertation, we conducted a retrospective analysis of urinary continence recovery after radical prostatectomy for individuals who completed the Michigan Urological Surgery Improvement Collaborative (MUSIC) - Patient Reported Outcomes (PRO) questionnaires. We assessed variability in urinary function outcomes at four post-operative time points and quantified the variability attributable to patients and surgeons (Aim 1). We trained pre- and post-operative prediction models that estimated long-term urinary continence recovery by predicting urinary domain scores and pad use at multiple time points (Aim 2). Lastly, we evaluated how temporal changes in practice patterns affect the robustness of models that estimate pad use by assessing various model building strategies (Aim 3).

Using MUSIC registry and patient reported outcomes data, we found wide variability in continence recovery, with greater variability attributable to patients than surgeons (66% versus 7%) in mixed-effects models. Models that incorporated post-operative data and predicted outcomes at proximal time points performed better than pre-operative models or models that predicted outcomes at distal time points. Model discrimination remained stable and models while model calibration showed some indication of over-estimation in later years, but no evidence of

calibration drift emerged. By conducting a retrospective study of statewide registry data to quantify variability in urinary continence, predicting long-term recovery, and assessing the robustness of prediction models to secular trends in registry data, our aim is to further the understanding of factors associated with urinary continence recovery and advance the science around prediction models using patient reported outcomes data. The results of these analyses fill gaps in our understanding of urinary continence recovery and further advance the goals of the Michigan Urological Surgery Improvement Collaborative to improve the quality of prostate cancer care in Michigan.

Keywords: Prostate cancer treatment, patient reported outcomes, urinary continence recovery, prediction models , model performance

Chapter 1 Introduction

Prostate cancer is a common and heterogenous disease. Over the past three decades, we have developed a greater understanding of this disease. However, uncertainties and controversies are widespread in prostate cancer care. Progress is slow, incremental, and contentious. Professional interests, individual experiences, and training influence experts to arrive at different conclusions despite seeing the same data. The tension between those who believe in the efficacy of screening and aggressive treatment versus those who argue for a more cautious approach is real in prostate cancer care. The purpose of this literature review is to understand how prostate cancer care has evolved over the past three decades and provide an overview of how patient engagement tools such as patient reported outcomes have been used in prostate cancer care. With uncertainties and controversies as backdrop, this review presents a synthesis of current literature while highlighting knowledge gaps in our understanding relevant to prostate cancer care.

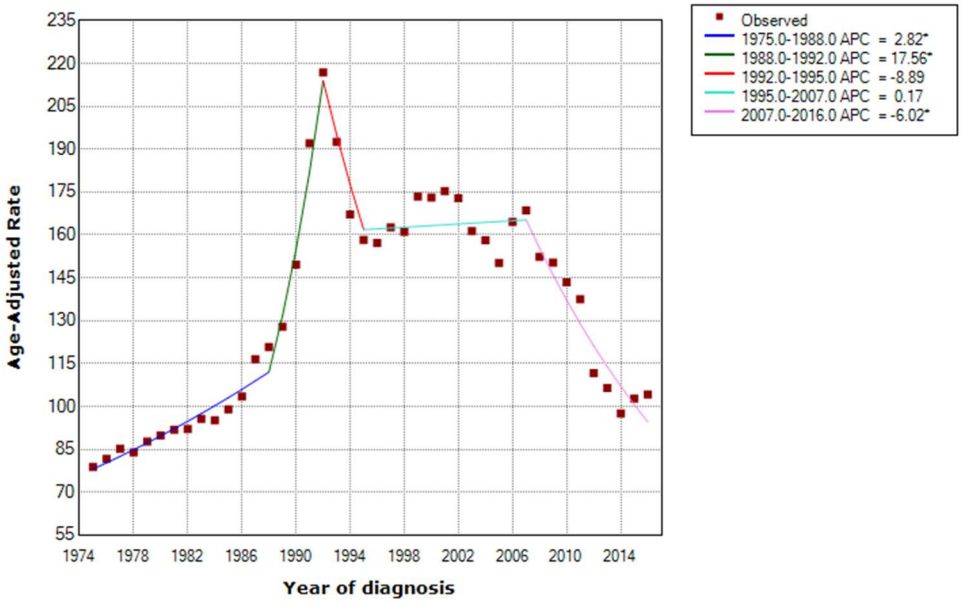
1.1 Epidemiology of Prostate Cancer

Prostate cancer is a common disease. In fact, it is one of the most prevalent malignancies in men. In the US, the lifetime risk of being diagnosed with prostate cancer is approximately 11.6% (95% C.I. 11.54-11.65), and the lifetime risk of dying of prostate cancer is 2.44% (95% C.I. 2.42-2.46).[1] An estimated 191,930 men will be diagnosed and about 33,330 will die of prostate cancer in 2020.[10] Prostate cancer incidence is highly sensitive to prostate cancer screening patterns. As prostate specific antigen (PSA) screening increased starting in 1991-1992, incidence rates increased as well.[3] In fact, a dramatic increase was observed in the incidence of prostate cancer between 1988 and 1992 by an average increase of 16.5% per year.[4] To fully appreciate the complexities in treatment and concerns about over-treatment, it is important to recognize how the prostate cancer landscape dramatically changed beginning in the mid-1980s with the arrival of the PSA screening era.

1.1.1 The Prostate Specific Antigen (PSA) Era

As PSA screening became widespread in the mid-1980s, we saw a rapid increase in prostate cancer diagnosis. The US National Cancer Institute's Surveillance Epidemiology, and End Results (SEER) data show that prostate cancer incidence rates increased rapidly between 1989 and 1992 (Figure 1) while the age-adjusted mortality rates continued to decline (Figure 2).

Figure 1. Prostate Cancer Incidence Rates (SEER 9). All Races, All Stages.



There are two primary reasons cited in the literature that contributed to the precipitous increase in incidence rates. Epidemiologists claim that the rising uptake of PSA testing contributed to this increase, including analyses of SEER and Medicare claims data, which have established that PSA testing led to the rapid increase in prostate cancer incidence.[5] Some researchers have speculated that the increased use of prostate biopsies might also have contributed to the increase in incidence rates.[6] However, a separate retrospective study of SEER and Medicare claims data conducted in 1998 found that transurethral resection of the prostate (TURP), a biopsy procedure that became popular during time, explained much of the observed increase in overall prostate cancer incidence between 1973 and 1986, however, the influence of TURP on the trend and overall magnitude of the rates diminished between 1987 and 1993.[6]

Suffice it to say, the rapid increase in PSA use and possibly increase in the use of TURP contributed to a rapid increase in prostate cancer incidence rates between 1988 and 1992. Notably, between 1992 and 1995, the incidence rates started to decline as a result of a *clearing out* of the prevalent cases[3]—i.e. reduction in the pool of indolent cancers that were initially undiagnosed.[4] The incidence rates plateaued between 1995 and 2007, but we did not see any further declines until after 2007 (Figure 1).[4] It is important to note that in these intervening years, United States Preventive Services Task Force (USPSTF) revised its screening guidelines in 2008 urging a cautious approach to PSA screening and in 2012, it gave a Grade D recommendation, advising against routine use of PSA screening and discouraging the use of Digital Rectal Examination (DRE) with both PSA and DRE were found to be an ineffective screening modality.[10,23]

The rapid increase in incidence in early 1990s brought about radical changes in stage and grade migration with an increasing number of men diagnosed with low risk, localized or indolent prostate cancer.[8] For example, in late 1990s to early 2000s, almost 80-90% of prostate cancer diagnosed was localized prostate cancer.[3,20] In contrast, before the PSA era, a larger proportion of prostate cancers were diagnosed at an advanced stage.[10] In addition to state migration, surveillance data suggest that PSA testing created age migration as evidenced by the changes in the age distribution of the population diagnosed.[3,11] For example, longitudinal analysis of SEER data reveals that prostate cancer is most frequently diagnosed in men between the ages of 65 and 74 with the median age of 66 at diagnosis.[12] Prior to the PSA era, the median age at diagnosis was 70 years.[3]

1.2 Prostate Cancer As a Health Policy Issue

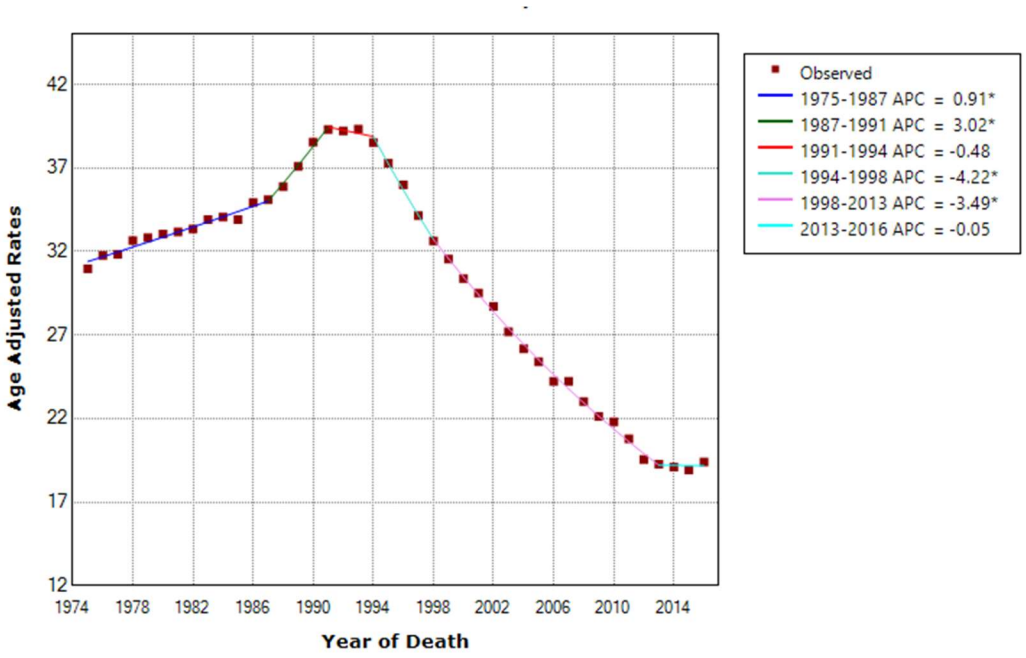
Prostate cancer prevention, screening, treatment, and survivorship issues have garnered remarkable interest in medical and health policy communities over the last several decades. As the US population ages, prostate cancer remains a significant public health issue with major cost, morbidity and mortality implications.[13,3,14] Annually, the National Institutes of Health invest over $200 million in studies to investigate prostate cancer prevention, screening, treatment and survivorship issues.[15] The National Cancer Institute estimates the national cost of prostate cancer care will continue to rise based on population estimates, which ranged from $11.85 billion annually in 2010 dollars to a projected $16.34 billion dollars in 2020 based on estimates of

cancer prevalence from 9 SEER areas.[13] As new and expensive treatment technologies are introduced, the overall cost of prostate cancer treatment is likely to increase substantially as new technologies are costlier than their traditional counterparts with uncertain health benefits.[16] Hence, covering the cost of prostate cancer treatment remains a significant health policy issue. As the nation grapples with the rising cost of health care and struggles to care for the aging population, the goals of providing the right care, for the right people, at the right time become even more important.

1.2.1 Controversies in Prostate Cancer Care

Prostate cancer is often an indolent disease in many men, and in some men the cancer is more aggressive and leads to death.[17] Surveillance data suggest that it remains the second leading cause of cancer death among men. However, there has been a decreasing trend in age-adjusted mortality rates over the past two decades with even more rapid mortality declines in recent years (Figure 2).[2,4,18] It is uncertain what caused the drop in prostate cancer mortality since 1991. Researchers hypothesized that it could be due to screening and treatment, changes in attribution of cause of death, and possibly due to increased risk of death from cardiovascular diseases among prostate cancer patients.[3] The increased risk of death from cardiovascular disease for patients diagnosed with prostate cancer is an important epidemiological phenomenon. As men get older, they are more likely to develop not only prostate cancer, but also are at a higher risk for cardiovascular and coronary artery diseases, diabetes, and metabolic syndrome among other chronic conditions. A large retrospective study of SEER data from 1973 to 2012 showed that prostate cancer patients make up one of the largest absolute numbers, which accounts for about 17%, of patients who have died of CVD.[19]

Figure 2. Prostate Cancer Mortality Rates. (SEER 9). All Races, All Stages.



* Indicates that the Annual Percent Change (APC) is significantly different from zero at the alpha = 0.05 level
Final Selected Model: 5 Joinpoints.

There is a lack of consensus on why mortality rates are declining. Clinicians who treat prostate cancer suggest that the observed improvements in mortality rates are due to better treatment options, including the introduction of robotic surgeries[20] and improving the precision of radiation therapies.[21] However, there is conflicting evidence indicating whether prostate cancer therapies truly prolong patients' lives pointing to evidence from a trial that compared surgery with active surveillance.[22] From an epidemiological perspective, this is a highly contested area with conflicting evidence leading to controversial conclusions.

Albert Mulley and Michael Barry, in their 1998 article published in the British Medical Journal (BMJ) calling for more high quality data in prostate cancer care, argued that "*the poorer the evidence, the more discretionary the interpretation, and the more controversial the conclusion*."[23] Since that article, dozens of large randomized trials in prostate cancer care attempted to provide higher quality evidence to answer lingering questions about screening and treatment efficacy. However, despite the increase in RCTs, challenges of poor evidence, discretionary interpretation, and controversial conclusions that Mulley and Barry identified remain large. Similar to the sentiments expressed by Mulley and Barry, Cooperberg et al. have argued that "*the absence of high-quality comparative effectiveness data, along with controversy*

regarding interpretation of the data that do exist, creates a fertile substrate in which variation would be expected to thrive."[24] Because of variations in treatment and concerns about over-treatment, in 2009, the National Academy of Medicine recommended examination of management strategies for localized prostate cancer on survival, recurrence, side effects, quality of life, and costs as one of the nation's most important comparative effectiveness research priorities to examine.[25]

1.2.2 Over-diagnosis and Over-treatment

Arguably, one of the most contentious aspects of prostate cancer care is the debate about overdiagnosis and over treatment. At the height of the PSA era, there was an increased concern that overdiagnosis of localized disease would lead to overtreatment of localized disease. However, a retrospective analysis of Medicare data for 2007-2012 shows that reduced PSA screening (post USPSTF revised recommendation) resulted in a 42% decline in population-based treatment rates, from 4.3 in 2007 to 2.5 in 2012 per 1,000 men compared to 8% decline (from 718 to 659) in treatment per 1,000 diagnosed men.[21] The investigators assert that decreasing rates of diagnosis, potentially from decreased screening, were the main driver for the decline in population-based treatment rates. This decline in prostate cancer screening and subsequent incidence is also reducing early stage prostate cancer diagnosis.[26] Interestingly, a smaller proportion of early stage diagnosis is resulting in diagnosis of more regional and distant-stage disease as a greater proportion of all prostate cancer diagnoses from 2010 to 2014.[4,27]

Evidence points to widespread over-treatment of prostate cancer that is detected from PSA screening, primarily when the disease is localized.[28,29,30,31] Researchers argue that had it not been for PSA screening, many of the cancers found during the PSA era would have never been diagnosed or diagnosed in later years.[3] Lack of early diagnosis of localized prostate cancer would have caused no clinical harm had they remained undetected.[10] However, once prostate cancer is diagnosed, there are several definitive treatment options available including radical prostatectomy, brachytherapy, external beam radiation therapy, proton beam therapy, androgen deprivation therapy in combination with other therapies, cryotherapy, high-intensity focused ultrasound (HIFU) among other options.[32] Non definitive therapies include active surveillance, watchful waiting and androgen deprivation therapy alone.[33] Morbidity associated with prostate cancer treatment remains substantial and no one treatment option is superior to other treatment

options, despite claims that technological advances have improved quality-of-life outcomes.[34,35,36] It is now widely recognized that most men die with prostate cancer than die from prostate cancer.[37] In an autopsy study of men who died of other causes, more than 20% of men aged 50 to 59 years and more than 33% of men aged 70 to 79 years were found to have prostate cancer.[10] Hence, over-treatment of prostate cancer remains a highly debated topic in prostate cancer care. Longitudinal trials suggest men who undergo treatment do not end up with better outcomes, both from a longevity and quality of life perspective than those choosing active surveillance.[22] Over-treatment of prostate cancer often leads to over-use of health care, creating burden on patients and care-givers. There is also evidence that suggests to over-treatment driven by specialty bias in which the specialists recommend the treatment options they perform.[38]

1.2.3 Specialty Bias

Prostate cancer are managed by urologists, urologic oncologists, radiation oncologists or medical oncologists.[33] Patients diagnosed with prostate cancer are generally presented with multiple treatment alternatives including whether active treatment or active surveillance is right for them based on their age, comorbidities, tumor grade, volume and other clinical characteristics. Patients are often given conflicting information by clinicians who specialize in different specialties, highlighting the need for multidisciplinary teams.[39] In some institutions, prostate cancers are managed by teams of multispecialty clinicians.[40] Traditionally, there has been a delineation of treatment modalities by specialty type. Urologists have been responsible for treatment of erectile and urinary aspects of prostate cancer care while focusing on surgical interventions. While medical oncologists typically focus on providing chemotherapy and radiation oncologists have generally undertaken radiation therapy.[33] A study conducted by Fowler et. al. in 1998, which included randomly sampled urologists and radiation oncologists, found that while these specialists generally agreed on the basic premise of screening and treatment, they overwhelmingly recommended the therapy that they themselves deliver.[41] A study conducted in Germany comparing treatment choices between urologists and oncologists, should they become patients with prostate cancer themselves, showed that urologists preferred surgeries while radiation oncologists preferred radiotherapy regardless of disease characteristics.[42]

There were many similarities between the German study and a US study, a survey of randomly drawn sample of over 700 urologists and radiation oncologists, asked these specialists to propose treatment based on a hypothetical case given for their clinical consideration. This study showed both specialties preferred the treatment modalities they offer, consistently chose primary treatments in favor of their specialty perceived the treatment they offered to be more effective and believed that their treatment would lead to a better quality of life.[43] Beyond surveys of specialists, meaningful physician-level variation in the management of low-risk prostate cancer has been documented in a robust study.[39] In a retrospective cohort study of SEER and Medicare claims data, Hoffman et al. found that patients whose diagnosis was made by urologists were more likely to receive upfront treatment, and when treated, more likely to receive a treatment favored by the urologists indicating the risk of overtreatment based on the specialist's personal biases.[39]

In light of these instances of documented specialty biases, provider selection becomes a critical part of decision-making for patients. Little is known about how patients with prostate cancer select providers. A small qualitative study conducted in Philadelphia area found that for screen-detected prostate cancer, the majority of patients relied on their primary care providers for referrals to diagnosing urologists, and on their diagnosing urologists to choose the treating specialist.[44] Hence, it becomes increasingly important for primary care providers and diagnosing providers to steer patients towards multidisciplinary treatment teams where patients are provided with balanced information and decisions are made based on the most robust evidence, patients' values, preferences and treatment goals.

1.3 Shared Decision-Making

Prostate cancer patients are encouraged to be more involved in their treatment decision-process.[45] The voices supporting shared decision-making have come from within the medical community. Elwyn et al. in their 2009 paper suggested that the unilateral imposition of professional opinion is no longer a valid mode of interaction in healthcare settings. They argued that medicine is undergoing a significant shift in how the roles of physicians and patients are defined, recognizing that decisions in medicine need to accommodate uncertainty that exists about the benefit versus harm ratio.[46] Perhaps in no other field than in prostate cancer, the trade-offs between benefits and harms is more relevant. There is a gradual push to accommodate

patients' preferences and values in screening and treatment decision-making to balance the trade-off between quality and longevity of life after a prostate cancer diagnosis. Nevertheless, clinicians who treat prostate cancer patients are challenged to present options in a balanced manner and often struggle to convey uncertainties and clinical equipoise to patients.[47,48]

Clinical equipoise is a relevant concept in prostate cancer care because of prominent clinical trials that have demonstrated that no one treatment option is superior to other in prolonging a man's life, and each treatment comes with a set of side-effects. Even though men are living longer with prostate cancer,[12] it is important to note that no one treatment option is considered to be the preferred method of treating localized prostate cancer. Hence, the balance, or equipoise, between the benefit and harm of various treatment options become relevant, making treatment decisions sensitive to the preferences of patients and providers.[46,49,90]

Preference sensitive care describes a situation where the evidence for the superiority of one treatment over another is not well established.[46] When a treatment is preference-sensitive, there is an increased emphasis on a collaborative process to arrive at a decision, jointly made by the patients and clinicians based on patients' values, preferences and treatment goals.[46] However, there is evidence to suggest that clinicians are not engaging in shared decision-making with their patients based on clinicians' self-reported lack of awareness and use of decision-aids, which often leads to patients receiving treatment without fully understanding the benefits and limitations.[50] This type of knowledge gap was demonstrated in a cross-sectional knowledge survey performed in newly diagnosed localized prostate cancer patients recruited through the Metropolitan Detroit Cancer Surveillance System in Michigan. Study results showed that both black and white patients had large knowledge gaps regarding the side effect profiles and survival benefits of different treatment options.[51] Given that knowledge gaps between side effect profiles and survival benefits of treatment options exist, this study raises particular concern that the preferences of clinicians likely superseded the preferences of patients highlighting the lack of true shared decision-making in prostate cancer care. Even though medicine might be moving away from paternalism to shared decision-making, the transition has been slow, difficult, and challenging in prostate cancer care as this review discusses in the subsequent sections.

To reduce unwarranted variations in prostate cancer care including concerns about over-treatment, the Michigan Urological Surgery Improvement Collaborative (MUSIC) has developed a roadmap for the management of favorable-risk, early-stage prostate cancer.[52] The roadmap

intends to *unlink* prostate cancer diagnosis from treatment through greater use of surveillance among other strategies.[53] The roadmap lists four steps in determining eligibility for active surveillance with the final step calling for clinicians to engage in shared-decision making to make a management decision either to pursue local treatment or active surveillance.[53,54] Additionally, to reinforce the clinical guidelines in the roadmap, MUSIC developed active surveillance appropriateness criteria and six quality measures to assess progress and provide feedback to clinicians on their management of men with low-risk prostate cancer.[52,53] These quality measures included consideration of active surveillance among eligible patients, confirmatory tests in patients eligible for active surveillance, and rates of verifiable/documented active surveillance among other measures.[54,55]

1.3.1 Risk Stratification and Treatment Decision-Making

Since clinicians may find it challenging to adequately risk-stratify newly diagnosed patients,[56] several organizations have put forward treatment guidelines to help clinicians identify ideal patients for active treatment or surveillance. These guidelines encourage clinicians to conduct confirmatory testing. Hawken et. al. conducted a study to test the sensitivity of contemporary active surveillance guidelines and assessed how many men who met active surveillance criteria proceeded with initial active surveillance.[57] The study concluded that urologists in Michigan had not coalesced around a single set of selection criteria for patient selection for active surveillance. Their study highlighted substantial differences in beliefs and perceptions around prostate cancer risk that effect treatment recommendations for patients with low risk tumors. The authors called for a better understanding of the entire decision-making process including characterization of provider perceptions of risk and thresholds for treatment, how cancer risk is communicated with patients, shared decision-making between patients and providers and the degree to which difference in patient preferences drive the variation observed in the analysis.[57]

1.4 Patient Reported Outcomes in Prostate Cancer Care

Historically, collecting information from patients about the effects of treatment was primarily accomplished through verbal history component of the clinician's assessments and symptoms checklists.[58] With advances in technology, there is a growing movement to collect information directly from patients in a structured format, commonly known as Patient Reported

Outcomes (PROs) or Patient Reported Outcome Measures (PROMs). These two terms are used interchangeably, however, PROs relate to any information on the outcomes of health care obtained directly from patients while PROMs are standardized or structured instruments or questionnaires that capture the status of a patient's health condition directly from the patient, without interpretation by health care professionals.[59,60,61]

As such, PROs are instruments designed to assess common issues that affect patients' health or quality of life after a diagnosis or treatment. Typically, patient responses to PRO questionnaires generate numerical scores, which reflect health-related quality of life and or health outcomes. PROs assess outcomes that only patients are able to provide based on how a given treatment is impacting their quality of life, functional status, symptoms, side-effects, and overall experience with care. PROs can be used for multiple purposes including efforts to improve patient outcomes and experience, to assess health care related quality of life and treatment associated morbidity, and to support treatment related shared decision-making process.[59] There are several types of PRO instruments; some are generic, condition or disease agnostic, that assess health related quality of life (HRQOL) while others are condition specific that assess outcomes related to specific conditions such as PRO instruments developed for oncologic care.[59,62] Traditionally, 5 different categories of PRO instruments are used in research and clinical practices for different purposes including; (a) health related quality of life (HRQOL) measures that can be generic or condition specific, (b) functional status measures that reflect patients ability to perform specific activities, (c) symptoms and symptom burden measures that are specific to type of symptom of interest, (d) health behaviors related instruments that specific to type of behaviors, and (e) patient experience measures that pertain to patient satisfaction with process of care delivery.[61]

PROs have traditionally been used in clinical trials and in comparative effectiveness research to evaluate treatment efficacy and effectiveness. In clinical trials, PROs have been used to assess how patients respond to different therapies, longitudinally track patients' health, and facilitate treatment modifications.[62,63] A systematic literature review of peer-reviewed articles detailing randomized controlled trials (RCT) published from January 2004 to March 2012, found that the trend to incorporate patient-reported data in clinical trials has been increasing over the past two decades.[64] Many RCTs include patient reported outcomes as useful and valid endpoints in addition to traditional clinical outcomes such as morbidity and mortality.[65] Since quality of

life is often the primary end-point for many cancer treatments that emphasize symptom management and palliation, PROs are increasingly being used to determine success of cancer treatments developed to improve patient's quality of life or treatments with palliative intent.[66,67] As such, the Food and Drug Administration (FDA) plays an influential role in defining patient reported outcomes, how and when they should be used. In doing so, the FDA has provided extensive guidance on how the Administration evaluates evidence generated from PROs. The FDA advises PRO instruments to be used when measuring a concept best known by the patient or best measured from the patient perspective.[60] In addition to being used in clinical trials, collecting treatment side-effect information directly from patients in a systematic manner are increasingly being implemented in routine clinical practice for quality improvement, performance measurement and to enhance patient-centered-care.[57,61]

1.4.1 PROs to Enhance Patient-Centered Care

In recent years, we have seen a greater emphasis on integrating PROs in routine clinical practice and in settings more convenient to patients to enhance patient-centered care.[58,62] Understanding and incorporating patient satisfaction and outcomes that include the patient's perspective are considered to be integral to a patient-centered care.[68] Some scholars have argued that patient-reported outcomes should be accessible in a clinically meaningful context to patients and providers in a manner that is similar to how laboratory results are routinely available and examined in clinical care.[69] PROs are viewed as instruments that could fill in the gap in measuring outcomes by focusing on items that are most relevant and important to patients. Basch et al. found that integrating patient-reported outcomes into clinical practice improves symptom control, communication and patient satisfaction and in absence of patient reporting of symptoms, clinicians tend to miss or underestimate many of the symptoms experienced by patients.[70] In addition to promoting patient-centered care, PROs have been promoted as quality improvement or quality assurance tools by delivery systems, health insurers, and increasingly the Centers for Medicare and Medicaid Services (CMS).

1.4.2 PROs for Prostate Cancer Quality Improvement & Value-Based Payments

Since PROs are useful in monitoring performance and stimulating quality improvement at the provider level, there is an increasing push by payers and government entities to include them in quality measurement as well as in value-based payments.[66] The Centers for Medicare

and Medicaid Services (CMS) is advancing patient reported outcomes in the context of value-based payment reform despite the fact that the current oncology performance measures advanced by the CMS do not incorporate PROs.[68,71] In late 2019, the CMS has released details of a proposed alternative payment model for oncology care, which includes gradual implementation of patient reported outcomes for symptom monitoring.[71,72] Likewise, several national organizations including the National Quality Forum (NQF) have endorsed of the use of PRO-based performance measures (PRO-PM) for the purposes of performance improvement and accountability.[73] PRO-PMs are patient reported outcomes that are primarily designed for performance improvement and accountability purposes, where patient generated data are aggregated for a hospital or a clinical practice.[70] As value-based and alternative payment models gain prominence, PRO-PMs are likely to be extensively used in performance improvement and accountability.

Even though, patient-reported outcomes can serve as quality improvement tools at the provider level as well as tools to improve outcomes at the patient level, a growing body of literature raises concerns about the use of PRO-PM in quality measurement. Specifically, a debate in clinical quality improvement involves whether adherence to process measures results in better patient-centered outcomes. Sohn et al. conducted a study designed to determine whether adherence to nationally endorsed quality measures was associated with patient reported functional outcomes including, patient satisfaction and treatment related complications. The study found weak associations between compliance with nationally-endorsed quality measures and patient-centered outcomes.[74]

Part of the reason why compliance with nationally-endorsed quality measures fail to improve patient outcomes is because most of the quality measures developed in the United States follow Avedis Donabedian's structure, process and outcomes model, which serves as the theoretical framework for quality improvement.[68] Most quality measures, including those related to prostate cancer, are primarily structure or process measures and only a few are outcome measures that are relevant to patients.[74] For example, Gori et al. conducted a structured review of contemporary literature in 2017 to identify all prostate cancer quality measures proposed by the scientific or the measurement communities (e.g. the National Quality Forum) and assessed the frequency of these quality measures appearing in peer-reviewed publications.[75] Their study revealed that out of 71 proposed quality measures, process measures accounted for almost 84.5%

(n=60) of all proposed measures with 7% (n=5) structure measures and only 8% (n=6) were outcome measures. At the time of the publication of their paper, only 7 of these 71 quality measures were endorsed by the National Quality Forum, which influences how physicians are reimbursed.[75] This study further revealed that the least studied measures in the literature were related to patient outcomes including assessment of sexual function and urinary incontinence. Hence, critics have a grim view of performance measures as they fail to address the items most important to patients and measures that are clinically meaningful that are valued by patients, providers and payers. [75,76]

In addition to the question of whether quality measures lead to improvements in patient outcomes, there is also a debate about whether PROs lead to actual improvements in the quality of patient care.[77] It is widely recognized in quality improvement circles that patient satisfaction with outcome is the ultimate arbiter of quality, but most measures of satisfaction focus on the process of cancer care rather than the outcome of care.[78,79] For example, quality measures that assess whether providers offered counseling or treatment choices without providers seeking patients' preferences, it is unlikely that the quality measure would end up improving the outcome of care as patients may not follow-up with the intervention as providers failed to take patient preferences into account.[76] Kotronoulas et al. conducted a systematic review to explore whether inclusion of PROs in routine clinical practice is associated with improvements in patient outcomes, processes of care and health services outcomes. They found that the routine use of PROs produced effect sizes that were predominantly small-to-moderate indicating tentative evidence regarding the effectiveness of PROs in improving quality of patient care.[77] The Kotronoulas study highlights the need to focus on outcomes measures to improve quality of prostate cancer care.

1.4.3 Common Prostate Cancer PRO instruments

As discussed above, treatments for prostate cancer includes active surveillance and active or definitive treatment such as radical prostatectomy, brachytherapy, radiation therapy among other options. Men are living longer after prostate cancer diagnosis and there is evidence to support that advances in treatment are reducing overall mortality despite longitudinal trial evidence indicating that a treatment option cannot be ruled superior to another treatment option. Each treatment option is associated with respective side effects impairing health-related quality

of life, which are becoming increasingly important.[64,80] Hence, prostate cancer specific PROs are deployed to help clinicians monitor treatment side effects and assess how patients progress on a given treatment over time. Literature suggests that urology specific PROs are appropriate tools to measure effects of prostate cancer treatment including urinary, sexual and bowel functions among other quality of life issues such as the ability to perform day to day activities that are most important to prostate cancer patients.[59,81,82]

Several patient reported outcomes measurement tools have been developed and some, but not all, have been tested and validated in urologic care including the University of California, Los Angeles Prostate Cancer Index (UCLA-PCI), Expanded Prostate Index Composite (EPIC-50) and the short-form (EPIC-26), International Index of Erectile Function (IIEF-5) – Sexual Health Inventory for Men (SHIM), International Prostate Symptom Score (IPSS), the Patient Reported Outcomes Measurement Information System (PROMIS) Interest in Sexual Activity and Satisfaction With Sex Life, The European Organization for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ), and the American Urological Association Symptom Index (AUA-SI) are more prominent PROMs.[83-84]

In prostate cancer care, the most commonly used PRO instruments are UCLA-PCI, EPIC-52/EPIC-26, and EORTC QLQ-C30.[67] The 20-item UCLA-PCI was one of the first PROs focused on treatment side-effects of prostate cancer. It was validated with 255 respondents with 84% of whom had localized disease. The initial design of the UCLA-PCI PRO included patients and their spouses resulting in increased relevancy for patients making treatment decisions for localized disease.[83,84] UCLA-PCI includes six disease targeted domains measuring function and bother in the urinary, sexual, and bowel domains. It has been extensively validated in multiple patient populations and reflects good cross cultural validity. However, the UCLA-PCI was criticized for not capturing the full spectrum of quality of life changes after localized prostate cancer treatment resulting in the development of EPIC-50.[83]

EPIC was developed to address some critical shortcomings of existing UCLA-PCI. EPIC includes measures that are relevant to prostate cancer patients including irritative and obstructive urinary symptoms, hormonal symptoms, and quantifies function specific bother between urinary, sexual, bowel and hormonal domains.[85] EPIC was tested and validated in a cohort of 252 randomly selected localized prostate cancer patients with equal representation of subjects between brachytheraphy, external beam radiation, radical prostatectomy and hormonal therapy to

ensure side effects from each of these therapies are well represented.[85] EPIC is widely used in prostate cancer research and practice with high test-retest reliability, internal consistency and criterion validity etc.[74,83,85] EPIC assesses function by identifying how frequently a prostate cancer patient has been affected by a treatment side-effect during the past 4 weeks. EPIC also assesses how bowel, bladder, hormonal and sexual side-effects bother a prostate cancer patient by seeking to quantify how big a problem have these side effects been to the patient in the past 4 weeks.[86] EPIC became a robust tool as a result of continuous improvement of UCLA-PCI instrument in which the 50-item EPIC included 17 of the original UCLA-PCI items.[83]

However, with 50 items, EPIC was considered to be too long and time-consuming on patients and when combined with other patient-reported instruments it is even more onerous on patients.[69,83,87] Hence, a shortened, 26-item version was developed by Szymanski et al. to facilitate quality of life measurement in research and clinical settings by identifying items suitable for elimination while retaining items that could still help measure the 5 symptom domains of EPIC-50.[87] Similar to the original EPIC, each item in EPIC-26 is scored from 0 being the least favorable to 100 being the most favorable health-related quality of life or outcome satisfaction. [69,78,85] EPIC-26 has gained traction because it has been tested and validated and is the one of the most widely used instruments.[69] It has also received international recognition as the International Consortium for Health Outcomes Measurement (ICHOM) has given its support for its wider use over other PROs currently in existence.[88]

Recognizing EPIC-26 could still be too cumbersome to administer in routine clinical setting, Chang et al. further shortened EPIC-26 to develop EPIC-CP, a one-pager PRO instrument with 16 items for clinical practice, which they validated in a cross-sectional cohort of 307 prostate cancer patients.[89] There are many similarities between EPIC-26 an EPIC-CP, however, there are two key differences between these instruments. First, these two instruments are scored differently and the directionality of the scores are opposite of each other. In EPIC-26, items scores need to be transformed to 0-100 scale, however, in EPIC-CP, item scores are summed up to arrive at domain scores (0-12 for respective domains), making EPIC-CP user-friendly tool for busy clinical practices. Likewise, in EPIC-26, as explained earlier in this section, the higher the score, the better the HRQOL. However, in EPIC-CP, the lower the score, the better the HRQOL.[90] Developers of EPIC-CP argue that they designed this PRO instrument to bridge the gap between research and clinical settings but it is important to recognize that

shorter instruments may lack the breadth and granularity of expanded instruments. The difficult trade-off between what is practical and useful versus what is meaningful and comprehensive in clinical setting is quite real, leading to implementation barriers and challenges.

1.4.4 Implementation Challenges

Integrating PROs into routine clinical practices is increasing encouraged by the transition to value-based health care and a growing body of evidence demonstrates PRO's impact on patient outcomes.[59,82] However, there are challenges to PRO integration into routine clinical care including operationalization of PROs, burdens on patients, providers and systems.[68] The burden on patients is an important aspect to be considered prior to implementing new interventions, especially when data are sparse to support improve quality of care with the use of PROs.[66]

There is also a growing concern about whether clinics are using the right PRO instruments for right types of indications and if the selected questionnaires are validated for specific indications or validated for the population of interest.[71,68] Some PROs are developed to understand certain aspect of care and may not adequately or reliably measure all dimensions of patient's functions. For example, some prostate PROs focus more on urinary incontinence, while others focus on irritative and obstructive urinary symptoms.[67] Some PROs focus on sexual interest while others focus more on erectile functions making it difficult for clinicians and practices to identify the right instruments for the dimensions of care or side-effect profile of a treatment of interest. When PROs are not always used as intended, it may result in measurement error and proliferation of misleading or erroneous information, hence, decreasing confidence in study results.[83]

To help address these issues, a national level effort is in place to streamline, test, validate and help facilitate the implementation of PROs. The Patient Reported Outcomes Measure Information System (PROMIS) is a federally supported measurement system that advances the science of patient reported outcome measures by offering item banks of patient reported outcome surveys for adults and pediatric patients to be used in clinical and research settings.[73,76] PROMIS offers a set of psychometrically sound, patient-centered measures that evaluates and monitors physical, mental, and social health and are designed to enhance communication between patients and clinicians.[91] These core PROMIS domains are created to be relevant for the assessment of symptoms and functions in multiple contexts, conditions, and diseases as these items generally

do not carry attributions to a specific condition or treatment.[92] PROMIS measures are purported to be better than conventional measures for many reasons including having a larger range of measurement to decrease floor and ceiling effects, having fewer items than most conventional measures to decrease respondent burdens and most importantly providing a common metric for interpretation and crosswalks (linkage) between many conventional measures for comparatibility.[91]

1.4.5 Clinician Engagement

Another challenge around implementation is the usefulness of PRO reports. It is understood that providing PRO to individual surgeons helps them to counsel patients, improve surgical technique and guide follow-up care decisions,[93] however, it is unknown to what extent clinicians review their PRO reports or act on the data presented in these aggregated reports. In Michigan, the Michigan Urological Surgery Improvement Collaborative (MUSIC) is implementing MUSIC-PRO to improve quality of care after radical prostatectomy. Participating urologists encourage patients to complete MUSIC PRO questionnaires at baseline and in months 1, 3, 6, 12 and 24 post radical prostatectomy. MUSIC has seen a continuous growth in PRO enrollment, with an overall goal of enrolling 70% of all patients who chose radical prostatectomies annually in PRO. MUSIC met this goal in 2018 with around 70% enrollment[94] and surpassed it in 2019 with over 75% enrollment in MUSIC PRO.

Recent literature suggests that there are variations in PRO engagement among MUSIC surgeons with high PRO engagement was associated with better patient outcomes, particularly among lower volume surgeons.[95] To minimize variations and improve the quality of care, MUSIC has been focused on expanding PRO use among MUSIC affiliated urologists. MUSIC also supports many activities by providing a framework to identify surgeon characteristics or techniques that may be useful in improving functional outcomes through activities such as skills workshops, video review, and mentoring.[95] Similarly, MUSIC measures and reviews urologists' outcomes and engages them in continuous quality improvement efforts. Despite these ongoing quality improvement efforts, MUSIC has yet to achieve the collaborative-wide goals for functional outcome improvements after surgery.[96]

One of the key quality improvement goals for MUSIC PRO is to improve two patient related outcomes after radical prostatectomy. These two key outcomes are related to social

continence, which is defined as maintaining temporary use of 1 pad per day or fewer (rate of 0-1 pads per day) at 3-months post-radical prostatectomy.[96] If pad use is limited to 0-1 pad per day then it is understood as maintaining social continence. MUSIC's 3-month social continence goal is 75% in the first 3 months of radical prostatectomy. Statewide average of MUSIC enrolled patients indicates that currently 64% of patients are maintaining social continence, which is about 11% below MUSIC goal. Similarly, MUSIC's 6-month urinary function goal is at least 90% of patients maintaining social continence and as of the end of 2019, 83% MUSIC enrolled patients were maintaining social continence.[96] Statewide outcomes of social continence at 3 and 6 months post radical prostatectomies remain plateaued in Michigan for the past since late 2014. There might be a few explanations of why the social continence rates have remained unchanged in Michigan. Perhaps, perhaps there is a "ceiling effect" where further progress is unattainable or perhaps urologists have not a way to further improve their skills, or there are other important reasons that we have not fully understood.

1.4.6 Clinical Interpretation Problems

Clinical interpretation of PROs is increasingly gaining traction as clinicians struggle to interpret changes in domain scores in a clinically meaningful manner. More specifically, clinicians find it challenging to understand if a patient reported function score at baseline is better or worse than a score the patient reports after a treatment as most PROs combine questions together into domains and generate composite scores to give a summative view on patient functions.[90,90] The summative view on functional status as represented by a score is useful to examine trends or average changes in HRQOL but it can also present challenges in clinical interpretation especially when the sexual, bowel, hormonal, and urinary domain score thresholds that should be considered clinically relevant are not defined.[90] To help build evidence for what constitutes clinically meaningful change from pre-treatment to post-treatment PRO scores, Skolarus et al. conducted a study to establish a score threshold that constitutes a clinically relevant change, or minimally important difference (MID) for each urinary, sexual, bowel and hormonal domains of the EPIC-26 instrument.[69] MIDs refer to the smallest change in domain specific summary score, usually at or above a given threshold, reflecting a clinically meaningful or important HRQOL change that is perceived by a patient.[65,90] Skolarus et al. found that

clinically meaningful changes in EPIC-26 scores ranged from 4 to 12 points depending on the domain, which they argue provide useful endpoints for clinical trials, comparative effectiveness research and the clinical care of men with prostate cancer after treatment.[69]

The field of patient reported outcomes in prostate cancer care is still evolving. Very few PROMs are adequately validated in prostate cancer care. Dowrick et al. reviewed the psychometric properties of 6 most commonly used PROMs found that several, including SHIM and IPSS, have not undergone formal evaluation of their measurement properties in surgically treated localized prostate cancer patients.[83] Similarly, the Patient-Reported Outcomes Measurement Information System (PROMIS) Sexual Interest and Satisfaction single item measures in patient with prostate cancer has not been validated.[97] Using unvalidated instruments in routine clinical practice could contribute to measurement error, wasteful effort to more serious patient safety issues and potentially create a false sense of success.

1.5 Prediction Models

Prostate cancer is a heterogeneous disease with a range of potential oncologic and functional outcomes that make treatment decision-making a challenge. Effective management of prostate cancer requires accurate prediction of potential outcomes. Several nomograms, probability tables, and prediction models have been developed to predict outcomes of interest including prediction of pathological features, presence of extra-prostatic disease, lymph node involvement, seminal vesical invasion, risks of metastases, biochemical recurrence after definitive therapy, and disease free survival among other endpoints.[98–101] The Partin Tables, Kattan, Briganti, and the Memorial Sloan Kettering Cancer Center nomograms and probability tables widely used in prostate cancer care.

1.5.1 Static Models

A vast majority of existing prostate cancer nomograms, probability tables, and prediction models are static models lacking continuous monitoring and updating. Model performance and validation are generally considered a one-time activity.[102,103] Prostate cancer prediction models have traditionally relied on preoperative clinical and tumor related factors such as patient's age, BMI, Charlson comorbidity index, tumor stage, biopsy grade, gland volume, and PSA levels to predict outcomes of interest.[104] For instance, the Kattan nomogram predicts disease recurrence

post radical prostatectomy based on preoperative variables such as PSA, clinical stage and biopsy Gleason scores.[105] Similarly, the Partin tables were designed to predict pathological stage and other endpoints based on clinical stage, serum PSA and Gleason score.[106] Other Partin tables have also been developed to predict biochemical recurrence after radical prostatectomy and radiation therapy. Partin tables remain one of the most widely used prediction tools in prostate cancer and have gone through several revisions.[99,107]

Similarly, the prostate cancer prediction models at the Memorial Sloan Kettering Cancer Center (MSKCC) were originally developed to predict organ confined disease, extra-capsular extension, lymph node involvement and seminal vesical invasion. The MSKCC models use patient's age, PSA level, clinical stage, biopsy Gleason score and the number and percentage of the positive biopsy cores to build the prediction models.

Though the above discussed prediction tools have advanced the field of prediction modeling in prostate cancer care, they are not without limitations. For example, Gandaglia et al. have argued that clinical prediction models that are built on pre-operative variables should be complemented by including imaging and biomarkers for better cancer staging.[108] Others have echoed similar sentiments supporting the inclusion of multi-parametric MRI and novel biomarkers in model development.[109,110]

Recognizing the limitations of static models that rely on preoperative clinical variables, some prediction models have added dynamic modeling features and novel biomarkers to ensure models remained relevant by periodical updating.[109,111] Recent prediction models have also incorporated patient reported outcomes in addition to patient demographics, clinical and tumor specific data.[112] In recent years, some novel biomarkers have been incorporated in existing prediction models. For example, Kattan et al. have added molecular markers such as interleukin 6 (IL-6) with traditional clinical variables to develop nomograms that would predict BCR post radical prostatectomy.[98,113]

Another limitation of static models is the statistical approach used to develop them. The prediction tools discussed above are typically based on classical statistical methods, including multivariable logistic regression models or the Cox proportional hazards models, to identify subgroups who would do better or worse on a treatment regimen.[104,114] Limitations of traditional statistical approaches to handle time-varying predictors have been raised[115]; however, machine learning approaches are not immune from flaws, including over-fitting among other

limitations.[116–119] In response, dynamic models have incorporated more complex approaches, including advanced regression techniques, pattern recognition, classification, decision trees and other machine learning approaches in additional to classical regression techniques.[120] There have also been advances in methodologies as it pertains to dynamic models, including adding more emphasis (i.e., weight) to newer patients so they have greater influence on model coefficients than more recent patients have a greater influence on model coefficients than the historical cohort.[111]

1.5.2 Dynamic Models

As static prediction models are not routinely updated and validated, dynamic prediction models have been proposed to move away from one-time model development and validation, reflecting an evolving health care environment.[102,116] In fact, there has been a steady growth in dynamic models in more recent years, which are designed to evolve over time as data on new individuals become available.[115,121] For example, Vickers et al. (2017) discussed their experience developing dynamic models at MSKCC, describing how their static models are transitioning into dynamic models that resemble a Netflix algorithm.[111]

Dynamic models require a continual flow of data with Electronic Health Record (EHR) and disease registry data are well-suited for these models as opposed to epidemiological studies or clinical trials. Prediction models based on routine clinical practice data representing the "real world" practice are more conducive to dynamic modeling as the flow of data is uninterrupted. Also, EHR data can provide time dependent/time variant predictors for the dynamic models to continually update.[104] Additionally, disease registry data collect information from diverse patients and clinical settings are also useful for dynamic model development as new data from new patients become available routinely from participating sites, hospitals, or geographic regions. In both instances, dynamic models can incorporate time varying data from existing patients as well as data from new patients.

1.5.3 Challenges: Dataset shift

It is known that clinical prediction models deteriorate over time for a variety of reasons including changes in technology, data structures, practice patterns, and patient population etc.

that occur in routine clinical care.[122,123] Various terminologies have been used to describe these changes including concept shift, concept drift, covariate shift, data fracture and dataset shift.[124] The said terms have been used inconsistently to describe the causes and/or the reasons for the deterioration in the performance of prediction models. For instance, different terms are often used to define the same problem, or the same terms are used to describe different problems.[124] Responding to the lack of consistency in terminologies and definitions, Moreno-Torres et al. provided a unifying framework and offered a standard term, dataset shift, to refer to the changes in underlying data.[124] Dataset shift is a general term given to describe the phenomenon when the data available at model development are not representative of the data on which the models are tested or deployed.[125] More specifically, Moreno-Torres et al. have defined dataset shift as a phenomenon that appears when training and test joint distributions are different.[124] There are several reasons for the differences in the joint distributions with sample selection bias and non-stationary environments being two primary reasons. Both of these concepts relate to changes in the distribution of underlying data. Sample selection bias relates to the introduction of bias in the selection of the training set. Selection biases in training set may cause dataset shift when applied in different testing environments. Non-stationary environments contribute to dataset shift when physical or temporal differences between training and test data sources occur.[124] Calibration drift is term used to describe the deterioration in model calibration as a result of temporal changes in underlying data. Temporal changes in clinical practice could silently alter the performance of prediction models, specifically impacting model calibration. This phenomenon is called calibration drift, a consequence of deploying models in non-stationary environments where differences arise in underlying data between the population on which a model was developed and the population to which the model is applied.

In addition to models becoming outdated with the passage of time, the models may not always represent the "real-world," limiting their usefulness. Many of the prediction models in urology were originally developed at single, high volume, tertiary care centers that may lack representation from patients served in community-based settings. Models developed in high volume, high acuity settings may not be ideal for low-volume, low risk settings as the models may over-predict an outcome.[112,126] Conversely, models developed in community-based settings might under-predict outcomes in high volume, high acuity settings. Hence, it is important to consider who were included (people) in model development, where (place) and when (time) were

the models developed, and if the original models have been retrained, validated, and recalibrated in other populations.

1.5.4 Dynamic Models and the Learning Health System

Dynamic models are well aligned with the concept of a learning health system. While dynamic models are about continuous monitoring and updating, Jenkins et al. have argued that continuous monitoring of model performance is inadequate.[121] The primary reason is the "data action latency," which is the lag between the availability of data and an appropriate action called for by the data. Solving the data action latency requires a mechanism to transport the results of continuous monitoring into the model to create a feedback loop for models to learn, retrain, and provide more accurate predictions. In an LHS, a system improves by learning from new data through cyclical process, creating new knowledge and using the knowledge to make further improvements. An LHS provides an infrastructure for dynamic models to be operationalized, similar to the idea of a learning prediction system envisioned by Jenkins et al. In a learning prediction system, clinical prediction models help improve a system by learning from the data, continually and in real time through cyclical learning loops.[102] The notion of a dynamic prediction model aligns with the virtuous cycle described by Friedman et al. in their landmark articles.[127,128]

Chapter 2 Dissertation Research Proposal

Problem Statement

Large clinical trials, cohort and retrospective studies have shaped our understanding of prostate cancer treatment outcomes.[129–133] However, progress have been incremental and often contentious as diversity of interests, experiences, and training often lead experts to arrive at different conclusions about the nature of the evidence. Learning *which treatment* leads to better patient outcomes in prostate cancer care requires an understanding of *for whom* and under *what* circumstances. Part of the *for whom* does a treatment work better question is increasingly being answered with complex statistical and predictive modeling techniques. Machine learning and statistical computing are accelerating how we investigate registry, clinical practice, and patient reported outcomes data to gain new insights and strengthen the evidence base. The objective of this dissertation is to use clinical registry and patient reported continence outcomes to assess heterogeneity in urinary continence recovery following radical prostatectomy, characterize the heterogeneity through prediction modeling, and assess the robustness of the findings to practice pattern changes.

The long-term goal of this work is to support successful implementation of prediction models developed with clinical registry and patient reported outcomes data into routine clinical practice. The development, deployment, and adaptation of prediction models in routine urology practices has the potential to support treatment decision-making, advance precision medicine in prostate cancer care and improve the overall quality of care. To achieve the objective of this dissertation, we propose to conduct a retrospective study of Michigan Urological Surgery Improvement Collaborative (MUSIC) registry, which includes the Patient Reported Outcomes (PRO) dataset to assess heterogeneity in continence recovery, dynamically predict continence outcomes and evaluate how prediction model performance is affected by underlying temporal changes in practice patterns. The overall research question this dissertation intends to answer is, "what are the patterns of urinary continence recovery, to what extent continence recovery is

predictable and how do changes in practice patterns affect the performance of prediction models over time?"

To answer the overall research question, we propose the following three specific aims:

Specific Aims

Aim 1: To assess heterogeneity in the recovery of urinary function following radical prostatectomy in a statewide collaborative. How much do surgeon and patient factors matter? Using linear mixed-effects models, we propose to assess heterogeneity in urinary function recovery after radical prostatectomy for patients who responded to the MUSIC PRO surveys. Mixed-effects models are robust regression models that extend classical regression models by considering the randomness or variability within and across patients, surgeons, and other grouping factors simultaneously and are appropriate for longitudinal (repeated measures) and nested (hierarchical) data. Our aim is to assess the patterns of urinary function recovery by constructing longitudinal mixed-effects models using patient demographics and clinicopathological characteristics and surgeon information. We will compare Expanded Prostate Cancer Index- Composite (EPIC-26) continence scores across baseline (preoperative) and post-operatively at months 3, 6, 12 and 24. *Hypothesis:* We hypothesize that there will be distinct continence recovery patterns that differ across surgeons.

Aim 2: Dynamically predict the long-term recovery of urinary continence following radical prostatectomy using preoperative and postoperative data. We propose to build dynamic prediction models that estimate long-term urinary continence recovery at different postoperative time points. Unlike static prediction models that are trained on baseline data gathered pre-operatively, dynamic prediction models update predictions at different time points as new data become available over the course of a patient's survivorship. Dynamic prediction models estimate the probability of continence over a time horizon as new data become available. Since a

dynamic prediction model uses time-varying postoperative data, it incorporates the 3-month PRO data and predicts outcomes at 6, 12 or 24 months post RP. Similarly, the 6 months model uses available data up to that point and estimates outcomes at 12 or 24 months. *Hypothesis:* We hypothesize that predictions made post-operatively will be increasingly accurate as compared to preoperative predictions.

Aim 3: To assess how temporal changes in practice patterns affect the performance of prediction models developed from disease registry data Performance issues related to prediction models have been widely documented in urologic care, primarily nomograms and other models predicting pathologic and other outcomes. Changes in practice patterns have shown to affect performance of prediction models. We aim to build prediction models using MUSIC registry and PRO data to predict patients' likelihood of achieving continence at 3, 6, and 12 months post RP and evaluate the performance of the model trained on 2017 data across subsequent years (2018 to 2021). By doing so, we aim to evaluate the magnitude of outdated risk probabilities to assess if models deteriorate over time as practice patterns change. Model deterioration has been recognized as a critical factor influencing the safety and successful implementation of prediction models in clinical settings.[134] As practice patterns change temporally, prediction models will systematically over- or under-predict risks as new data become available.

2.1 Specific Aim 1

To assess heterogeneity in the recovery of urinary function following radical prostatectomy in a statewide collaborative. How much do surgeon and patient factors matter?

We intend to assess patterns of urinary function recovery by constructing longitudinal hierarchical mixed effects models using patient demographics and clinicopathological characteristics and surgeon information to compare EPIC-26 continence scores across preoperative baseline and post-operatively at 3, 6, 12 and 24 months. Auffenberg et al. conducted a retrospective study to assess patient- and surgeon-specific factors associated with urinary outcomes 3 months post prostatectomy in MUSIC.[135] Our approach differs from Auffenberg's approach as they dichotomized surgeons into 'top performing' versus 'other MUSIC surgeons'

and assessed outcomes accordingly. The present study aims to assess heterogeneity in continence outcomes by building mixed-effects models to address the repeated, longitudinal, and hierarchical nature of MUSIC PRO data.

There are several reasons for the selection of mixed-effects models for our analyses. First, mixed-effects models are ideal for longitudinal analyses of repeated measures as PROs are collected at multiple time points. Longitudinal analyses of continence trajectories better reflect the recovery process than cross-sectional studies as recovery is a gradual process that occurs over many months. Cross-sectional studies provide a useful snapshot in time but are not ideal to model individual change over time. Further, mixed-effects models can handle uneven spacing of PRO data well and do not require the same number of PRO responses per patient in the study. These properties allow a patient who may have submitted PROs at months 3 and 12 but may not submit at months 6 or 24 to be included in the study.

Second, mixed-effects models are also appropriate for hierarchical, correlated, or clustered data as patients in this study are clustered within a surgeons and, hence, are likely to be similar. In mixed-effects models, clustered or correlated data are handled without violating the independence assumption. Third, mixed-effects models allow for the analyses of both fixed and random effects. Random effects produce the average effects or estimates of patient- and surgeon-level variability in outcomes that may be more generalizable to patients and surgeons from participating urology practices in Michigan. Mixed effects models quantify the amount of variability in urinary function recovery that can be attributed to patients and providers.[136,137] Fourth, in mixed effects models, time can be continuous or categorical (i.e., fixed set of time points). In our analyses, we propose to produce and compare estimates of continence recovery by considering time both as categorical and continuous variable. These properties make mixed-effects models flexible to analyze PRO data that are unevenly spaced repeated measures that are multi-level, where repeated measures are clustered within patients and patients are clustered within operating surgeons.

The mixed-effects models in this study will have 3 levels. Level 1 consists of repeated measures of PRO data. Level 2 consists of patients who reported their outcomes after radical prostatectomy. Level 3 consists of urologists who performed the surgeries. The overall structure

of the models is PROs reported at multiple time points nested within patients, and patients nested within urologists.

2.2 Methods

2.2.1 Data Source

We will use MUSIC registry data for our analyses. MUSIC maintains a registry of newly diagnosed prostate cancer patients in Michigan. The registry consists of 90% of state's prostate cancer patients from 46 community, private, and academic urology practices. The registry is established for quality improvement and collects data submitted by participating urology practices. Trained data abstractors prospectively enter a standardized set of demographic and clinicopathological data, including tumor characteristics, comorbidities, and treatment modalities among other relevant information into a web-based clinical registry.[138,139] To participate in MUSIC, each urology practice must obtain exemption approval from their local institutional review boards, as MUSIC efforts are intended for quality improvement.

2.2.2 Inclusion and Exclusion Criteria

Patients diagnosed with prostate cancer between 2014 and 2021 will be included in the analyses. We will exclude any cases of previous diagnosis of prostate cancer. Other inclusion and exclusion criteria will be determined during data analysis.

2.2.3 Data Analyses

To assess heterogeneity in continence recovery, we will first evaluate patient specific continence recovery for 3-month, 6-month, 12-month, and 24-months using descriptive measures. The distribution of patient specific continence recovery measures will be evaluated for each time point during study follow-up, including assessments of mean, standard deviation, median, and interquartile ranges. From these assessments we will be able to assess model-free or unadjusted patterns of urinary function recovery over time. Additionally, we will assess the nature and the degree of missing data. Patient-level graphs, such as spaghetti plots showing

individual trajectories, will be created as initial steps towards understanding the variations in trajectories. For provider level data, we will create quantiles based on surgeon's annual caseload to characterize the variations in urinary function recovery by surgeon volume. All appropriate data cleaning steps will also be implemented at this time. After completion of descriptive analyses, we will build models in a stepwise fashion that incorporate time trajectories. Model 1 will serve as our initial or unconditional means model, while the subsequent models will include time and other variables as predictors to assess temporality and heterogeneity in continence recovery.

2.2.4 Model 1: Random Intercept Only Model (Base Model)

We propose to build an intercept only model, with no predictors at any level. Patients and surgeons will be modeled as random effects. This model allows each patient and surgeon to have their own unique intercept, which will allow us to assess whether sufficient heterogeneity exists within-patients to support mixed-effects models. We will obtain the grand mean of the urinary domain scores across all patients the study. This model will serve as our base model and performance of subsequent models can be compared against this model.

2.2.5 Model 2: Unconditional growth and fixed-effects model

The linear or unconditional growth model introduces time as a predictor, which allows us to assess how much within patient variability in urinary function recovery can be attributed to patients while also determining how much variability we see at different time points (i.e., at months 3, 6, and 12). This model allows each individual patient to have a random intercept (if deemed appropriate from Model 1), and a random slope. Time will be entered into the model as a continuous variable to obtain one estimate of the mean slope across all measures; and, explored as a categorical variable to evaluate non-linearity of the slope across different time points.

2.2.6 Model 3: Full Model

In the full model, we will include predictors that are known to influence continence recovery including age, BMI, Gleason grade group, preoperative prostate specific antigen (PSA),

clinical T-stage, nerve sparing procedure, and surgeon's annual caseload. Similar to the intermediate model, time will be entered into the model as categorical and continuous predictors.

In all three models, we will compute the intraclass correlation coefficient (ICC) to estimate the proportion of variance in outcome explained by the patient and surgeon. The ICC calculates the ratio of within-patient variance by dividing the variance of the intercept by the sum product of the residual variance and the variance of the intercept. If the resulting ICC estimate is low (i.e., close to zero), then it would indicate that there is little correlation of the intercept within patients, and a fixed intercept may be a more appropriate model. If the resulting ICC estimate is high (i.e., close to 1), then it would indicate that a large proportion of the variability in the outcome can be attributed to the patient and allowing patients to have their own intercepts in the model is the appropriate choice.

2.3 Specific Aim 2

Dynamically predict the long-term recovery of urinary continence following radical prostatectomy using preoperative and postoperative data.

We propose to train random forest models to dynamically predict the probability of a patient recovering continence at specific postoperative time periods. Predicting continence is a challenge because it is highly affected not only by patient level factors (i.e., patient and tumor characteristics and variations in urethral length, etc.), but also by provider level factors (i.e., high/low volume, high/low performing surgeons). The goal of this approach is to assess if dynamic postoperative models perform better than static pre-operative models in estimating continence recovery.

The dynamic models could potentially support clinicians to accurately predict patient outcomes at different timepoints. By predicting outcomes at multiple time points, clinicians can assess how patients realize different recovery trends, which might add to the conventional understanding of patients who are more or less likely to recover early continence. A dynamic model may support treatment decision making and set realistic expectations of outcomes. The models we are proposing are appropriate for our research question and the type of outcomes that

we describe below as the similar approaches have been used in prior studies of patient reported outcomes in urologic care. [111,116,140–143]

2.4 Methods

2.4.1 Data Source

The Michigan Urological Surgery Improvement Collaborative (MUSIC) administers MUSIC-PRO questionnaires to improve quality of care after radical prostatectomy (RP). Participating urologists encourage patients to complete the questionnaires at baseline (preoperative) and at 3, 6, 12 and 24 months. We will include patients who completed MUSIC-PRO questionnaires between September 2016 and October 2021.

2.4.2 Study Design

Our aim is to train and validate dynamic predictive models to estimate the probability of continence recovery. Using patient demographic, clinical and patient reported outcomes variables, we will train our models to estimate urinary continence recovery months 3, 6, 12 and 24 months after radical prostatectomy. Studies have documented the difficulties in predicting continence after radical prostatectomy based on demographic, tumor, and surgeon characteristics.[17] Our models are intended to assess if dynamic prediction models trained with postoperative variables perform better than models trained with preoperative variables.

2.4.3 Cohort Construction

All patients who have undergone RP and completed MUSIC-PRO questionnaires at baseline, 3, 6, and 12 months post RP will be included in the study. The time frame of the study is September 2016 and October 2021. We will randomly split the cohort into derivation and validation cohorts. The derivation (training) cohort will consist of 66% of patients who completed MUSIC-PRO questionnaires at baseline and at 3, 6, and 12 months after prostatectomy. The validation cohort will consist of 33% of patients who completed MUSIC-

PRO questionnaires at baseline and other time points post RP. We will train prediction models on the derivation set and will evaluate their performance on the validation set.

2.4.4 Inclusion and Exclusion Criteria

Patients who completed a patient reported outcomes survey at baseline (pre-operative) and at 3, 6, or 12 months post RP will be included in the study. Patients younger than 41 years and older than 85 years at prostatectomy will be excluded. Patients who did not complete any post-operative MUSIC-PRO questionnaires (either 3, 6, or 12 months) but completed baseline MUSIC-PRO (pre-operative) will also be excluded from the study.

2.4.5 Random Forest Models

We propose to build two random forest models that separately predict binary (continence) and continuous outcome (urinary domain scores) at 3, 6 and 12 months after RP. A random forest (RF) model can be simply understood as a collection of decision logics in tree-like structures with leaves and nodes.[41] The RF model iteratively splits data into nodes and results are combined across trees.[13] RF models use a subset of predictors in each node of each tree through bootstrap sampling. Random forest is a versatile algorithm that can perform binary classifications or predict continuous outcomes.[42]

2.4.6 Predictors

To build our proposed models, we will use predictors that are routinely in the urology literature, including age, BMI, clinical-T stage, PSA, prostate volume, and baseline pad use, and continence scores. These predictors are reliable and widely used in studies that have predicted continence post RP. We will evaluate the relative importance of predictors, including ranking predictors by decrease in the Gini impurity index and summing the importance or influence of each predictor across all trees.

2.4.7 Coding and Cleaning of Predictor Variables

We will wrangle, clean, code, and transform data according to established protocols using "dplyr" and/or "tidyr" packages that are part of "tidyverse" in R. EPIC-26 scores will be transformed into a scale of 0-100 with 100 being the highest possible outcome based on scoring instructions established by Wei et. al.[52] Prior to running our models, we will create histograms and box and whisker plots to assess the distribution of study variables and identify outliers. We will assess frequency distribution for categorical predictors and will avoid categorizing any continuous variables to minimize information loss.

2.4.8 Outcomes

Our outcome is achievement of continence at 3, 6, and 12 months post RP. In the urologic literature, complete continence is defined as 0 pads per day, social continence is defined as 0-1 pad per day, and incontinence is defined as 2 or more pads per 24-hour period.[49] Our outcome is clinically meaningful as patients who undergo RP experience severe loss of urinary and sexual functions[50,51] and recovery of these functions are important not only to the clinicians but, more importantly, to the patients and caregivers.

2.4.9 Model Performance : Discrimination

We will evaluate the performance of our models in the validation cohort by assessing model discrimination and calibration. All of our models will be evaluated on the same test set to ensure fair comparison. We will use the area under the curve (AUC) on the receiver operating curve (ROC), one of the widely used diagnostic tools in machine learning. Higher AUC represents better model performance. AUC estimates the probability that a model correctly discriminates between individuals who are with or without an event or an outcome. Hence, ROC plots sensitivity (true positive rate) versus false positive rate (1-specifity). We will assess AUCROC at 3, 6, and 12 months post RP by predicting continence at these time points and comparing them with patients' true continence based on the PROs.

2.4.10 Model Performance: Calibration

We will use calibration plots to assess comparison between predicted values and observed values along the diagonal line, which provides a good sense of model calibration. Alternatively, we could assess model calibration by comparing observed versus predicted risks by risk groups. This method is based on a study by Ahmad et al. (2018), in which they grouped patients into deciles based on predicted risk before plotting observed versus predicted risks.[13]

2.5 Specific Aim 3

To assess how secular changes in practice patterns affect the performance of prediction models developed from disease registry data

We aim to train prediction models using MUSIC-PRO data to estimate the probability of continence recovery at 3, 6 and 12 months post radical prostatectomy. We plan to compare two model updating strategies against a default (baseline) approach. The model built under the baseline approach will not go through re-training while models build under the two updating strategies will be updated annually with different combinations of cohorts. The models will be validated in subsequent years (2018 to 2021). By doing so, we will assess if model performance deteriorates over time as practice patterns change. Deterioration of prediction models has been recognized as a critical factor influencing the clinical usefulness, performance, and successful implementation of prediction models in clinical settings using EHR data.[144,134] As practice patterns and/or patient populations change temporally, estimates of risk probabilities based on registry data from earlier years may over or under predict risks depending on the underlying changes in the registry data. The objective of this analysis is to assess if models trained on registry data deteriorate over time.

2.5.1 Deterioration of Prediction Models

Poor performance resulting from calibration issues can make predictions misleading and often harmful.[126] Model performance deteriorates when models are applied in a novel setting or in separate cohorts/populations. Performance may also deteriorate due to temporal changes as models trained on older data are less likely to remain robust over time. Similarly, abrupt changes in clinical practice, introduction of new technology or procedures, changes in reimbursement

policies, or changes in data collection or data definition could also deteriorate model calibration.[144] Methodologists have proposed that model calibration is not an inherent property of a prediction tool; but is a joint property of a model and the cohort to which the model is applied.[145] Hence, if a model is "well-calibrated" in one dataset, it does not mean the model is "well-calibrated" in another dataset, or in a different cohort or at a different time. Further, models may also experience calibration drift, which is changes in model performance that occurs gradually and often silently.

Calibration drift has emerged as a critical aspect of model performance in streaming data generated from electronic health records.[144] However, calibration drift in clinical registry data is not widely explored. Using the MUSIC registry, we propose to assess calibration drift, resulting from temporal changes in clinical registry data. We are interested in assessing how models may systematically over or under predict risks when the underlying data environment changes. Our intent is to characterize and estimate the magnitude of calibration drift in static registry data, which are being used for many purposes including development of prediction models and to assess provider performance.[16, 142]

2.5.2 Importance of Calibration in Prostate Cancer Prediction Models

Calibration is the measure of closeness of expected probability to the underlying probability (observed probability) of the population and relates to the reliability of risk predictions.[118] A model is considered well calibrated if, for every 100 patients given a prediction of 10%, the actual number of events is close to 10 or in probabilistic terms, the observed probability and expected probability are as close to each other as possible.[145–148] Calibration is as important as discrimination for a model to be clinically useful but it has been largely ignored in favor of measures that assess discrimination such as sensitivity, specificity, and AUC.[146,147,149] In prediction modeling, it is understood that discrimination analyses are most relevant for classification tasks, whereas calibration is important for prognostic problems[6] with methodologists contending that calibration is the primary requirement of a prognostic predictive model.[149,150]

2.6 Methods

Data Source: MUSIC Registry

2.6.1 Secular Trends in MUSIC Registry

Temporal changes in MUSIC registry data will be assessed using Joinpoint linear regression, which detects statistically significant annual changes. Joinpoint regression program software was developed by the National Cancer Institute (NCI) to detect statistically significant changes in temporal trends in cancer surveillance research and assesses temporal trends by fitting the simplest model to describe the trend data.[151] The user specifies the number of minimum and maximum joinpoints starting with a straight line with 0 joinpoints and then adding more joinpoints to determine whether multiple connecting lines better describe the trend over time by detecting changes in trends.[151,152]These changes are expressed as annual percentage change (APC) which is used to compare year over year trends. We will use NCI's Joinpoint Trend Analysis Software (version 4.8.0.1) to perform temporal analyses.

2.6.2 Model Development

We propose to develop multiple logistic regression models trained on the MUSIC registry data to predict continence recovery at 3, 6, and 12 months post RP in 2017 and evaluate the performance of the 2017 model across subsequent years (2018-2021). We will assess model discrimination and calibration for each year and statistically and graphically assess how calibration might deteriorate over time because of temporal changes in registry data.

2.6.3 Assessing calibration across multiple years

There are no established calibration measurement statistic that is widely accepted[153] and calibration is often assessed by testing for lack of calibration.[154] We will report calibration slope

and intercept together as summary measures to assess calibration.[155] We will also construct calibration plots to complement summary statistics.[147] Calibration is plotted by regressing the expected probabilities against observed risks. In an ideal setting, the observed risks are as close to the expected risks, resulting in a diagonal line in the plot with slope of one and intercept zero.[126, 150] We will present summary statistics and calibration plots for every model year to assess calibration drift.

2.6.4 Clinical Implications

With rapidly expanding disease registry data, electronic health record capabilities and the move towards sophisticated statistical computing, it is reasonable to expect that future prostate cancer patients would benefit from more personalized probabilistic predictions. This advancement is likely to positively impact future patients, but the possibility of harm caused by prediction models cannot be minimized. Simulation studies have indicated that poorly calibrated models yield inaccurate risk prediction for patients.[146] In prostate cancer care, poorly calibrated models may do more harm than good. Therefore, as more models are being integrated into EHRs, updating strategies to maintain performance are becoming critical components of model implementations.[144,156] From a clinical perspective, a model that has good discrimination but is mis-calibrated would cause a myriad of potential problems that could negatively impact patient care, clinician buy-in, contribute to higher cost, and negatively impact patient outcomes.[157]

Chapter 3 Heterogeneity in the Recovery of Urinary Function Following Radical Prostatectomy in a Statewide Collaborative. How Much do Surgeon and Patient Factors Matter?

3.1 Abstract

Objective: The objective is to characterize variation in urinary function recovery post RP in a statewide clinical quality improvement collaborative by assessing the amount of variability explained by patients and surgeons.

Methods: Using the Expanded Prostate Cancer Index Composite -26 (EPIC-26) survey and the prostate cancer registry data maintained by the Michigan Urological Surgery Improvement Collaborative (MUSIC), we retrospectively assessed variability in urinary function recovery among patients undergoing radical prostatectomy. We built longitudinal mixed-effects models to assess post-operative outcomes at 3-, 6-, 12- and 24-months. Models included surgeons and patients as random effects as well as demographic and oncological characteristics as fixed effects. Primary outcome is patient-reported urinary incontinence domain score after RP. Urinary incontinence domain score is a composite score that includes pad use and other urinary function items that contribute to incontinence severity. We calculated the EPIC-26 urinary continence domain scores, which range from 0-100 with higher scores reflecting better urinary function.

Results: We identified 9,159 men who underwent RP and completed PRO surveys at different timepoints during the study period. Mean (SD) urinary domain score at baseline was 90 (14.8), which decreased to 52 (27.4) at 3 months and increased to 67 (26.1) at 6 months. Descriptively, when stratified by surgeon's annualized RP volume, mean urinary continence was higher for patients of high volume surgeons 52.1 (27.4) versus 44.1 (26.8) for low volume surgeons. In the multivariable model, 73% of variability in continence was attributable to patients and 6% of variability to surgeons.

Conclusion: Considerable variability in urinary function recovery was observed. Variability attributed to patients was greater than the variability attributable to surgeons. In multivariable models, surgeon's annualized case volume did not fully explain the variability in urinary function recovery. Quantifying variability attributable to patients and surgeons may help to better understand the nature of urinary function recovery from patients' perspective.

3.2 Introduction

Urinary incontinence is one of the most bothersome adverse events after radical prostatectomy and the recovery is highly variable.[158–162] Urinary incontinence is often assessed with a single item pad use, while urinary function is assessed with a composite score that includes both pad use and other urinary factors that represent the spectrum of incontinence severity.[163–166] Different approaches have been used to quantify variability in incontinence and urinary function recovery. Some studies have assessed heterogeneity in incontinence rates between surgeons at cross-sectional time points while others have stratified surgeons by oncologic outcomes to assess if patients of surgeons with superior oncologic outcomes also have early continence recovery.[163–165,167,168] Assessing the amount of variability in continence and urinary function recovery is important for quality measurement, value-based reimbursement, and to improve patient outcomes over time. From a quality improvement perspective, variability attributed to known factors, such as surgical skills and training, are actionable[169,52] but an important challenge is to reliably quantify the magnitude of variability that is attributable to surgeons.

Wide variability in clinical and oncologic outcomes between surgeons are documented using clinician reported outcomes[162–164,170] with studies establishing the association between surgeon volume (case load) and clinical and oncologic outcomes.[162,170–178] Higher surgeon volume may also lead to better clinical and oncologic outcomes; however, the evidence is mixed to support the association between surgeon volume and patient-reported urinary function outcomes. Further, the existing studies are cross-sectional and do not adequately capture the longitudinal nature of urinary function recovery from patients' perspective. Though studies have shown both surgeon and patient factors impact urinary function recovery,[161,168,171,179,180] what remain unexplored are the amount of variability attributable to both patients and surgeons, and the

association between surgeon volume and patient-reported urinary function outcomes using PRO data.

Efforts to include PROs in the assessment of continence and urinary function recovery is still in its infancy. Higher surgeon volume may also lead to better clinical and oncologic outcomes, but studies have not conclusively established a volume-functional outcome relationship. The primary aim is to quantify variability in urinary function recovery by constructing longitudinal mixed-effects models compare urinary function scores preoperatively at baseline and post-operatively at months 3, 6, 12 and 24. The secondary aim is to explore the association between surgeon volume and urinary function recovery. We hypothesize that there will be distinct urinary function recovery patterns that differ across surgeons and patients at different time points.

3.3 Methods

This is a retrospective study of a prospectively maintained patient reported outcomes data to assess variability in urinary function recovery after radical prostatectomy. The Michigan Urological Surgery Improvement Collaborative (MUSIC) maintains a registry and administers patient reported outcomes surveys at multiple time points. MUSIC registry captures approximately 90% of all newly diagnosed prostate cancer cases in the state of Michigan and is supported by abstractors who enter data in the registry by searching the Electronic Health Records (EHRs) at their respective practices for prostate cancer cases. The patient data are entered and updated periodically as new data become available. Abstractors are trained by MUSIC coordinating center staff and use a standardized data entry protocol to enter cases into MUSIC's web-based clinical registry. Men diagnosed with prostate cancer who are considering radical prostatectomy are asked to complete baseline PRO survey which includes items from the Expanded Prostate Cancer Index Composite – Short Form (EPIC-26).[87] Briefly, EPIC-26 is a validated instrument that collects outcomes directly from men who underwent prostate cancer treatment. The study population described below includes newly diagnosed patients from MUSIC affiliated practices who underwent radical prostatectomy and completed the EPIC-26 survey.

3.3.1 Study Population

To assess the heterogeneity in continence outcomes, we identified men with newly diagnosed prostate cancer from January 2014 to October 2021, who underwent radical prostatectomy and responded to at least one PRO survey. Men who did not complete any PRO questionnaires and men who were previously diagnosed with prostate cancer were excluded. Similarly, men who were on active surveillance, watchful waiting, or received non-surgical treatments were not included.

3.3.2 Study Variables

Outcome variable: The primary outcome is patient-reported urinary function domain score following radical prostatectomy based on the responses from the EPIC-26 survey. Ordinal level responses to the five EPIC-26 urinary domain questions were transformed into a composite score between 0 to 100 with higher scores indicating better urinary function.[85]

Independent variables: The independent variables included baseline urinary continence score, patient age at diagnosis, pre-operative BMI (kg/m^2), Charlson comorbidity index, prostate specific antigen (PSA) test, Gleason grade group, clinical T-stage, and the highest percentage cancer involvement on a positive core. Surgeon characteristics included surgeon's annualized radical prostatectomy volume, calculated as the ratio between total number of prostatectomies performed by a surgeon over a duration, expressed in years; and whether bilateral nerve sparing technique was involved. Nerve sparing was dichotomized into bilateral versus none with partial nerve spare considered as an incomplete nerve sparing procedure.

3.3.3 Statistical Analysis

We assessed the distribution of patient level factors by comparing clinical, demographic, and oncological characteristics of patients who underwent radical prostatectomy . We examined urinary domain scores at baseline and at 3, 6, 12 and 24 months post-surgery stratified by

surgeon's annualized radical prostatectomy volume. Using the lme4 package in R statistical software (R Consortium, Vienna, Austria), we built multiple mixed effects models to assess longitudinal patient reported outcomes .[181,182] Consistent with the terminology of mixed-effect modeling, longitudinal measures of urinary continence were nested within individual patients to account for non-independence in repeated measures.[183] Time-invariant patient characteristics (assessed pre-operatively at baseline) were included as patient-level variables. Individual patients were nested within their respective operating urologist. Conceptually, the overall structure of the models can be summarized as repeated measures of continence nested within patients who were nested within urologists. By nesting repeated measures within patients and patients within surgeons, we sought to understand the degree to which measures were correlated within patients and surgeons. Patient characteristics, baseline continence score, and surgeon's annualized radical prostatectomy volume were included as fixed effects.

To assess the variance in continence outcomes, we fit 3 mixed-effects models[183] with post-surgical urinary continence as the outcome. The base (null) model included both patients and surgeons as random effects without independent variables allowing the null model to quantify urinary function recovery across all time points. The null model served as the unconditional means model that assesses the between-patient and surgeon variability in urinary function recovery. The intermediate model included time (6-,12-, 24-months), modeled as a categorical predictor. In the final model, we assessed the changes in variability as we introduced predictors such as patient age, BMI, Charlson comorbidity index and tumor characteristics, and baseline continence score in the fixed-effect structure, while keeping categorical time as a predictor and the random-effects structure consistent for comparison purposes. For each of the models, we calculated the proportion of the total explained variability at the patient and surgeon levels and assessed how variability at the patient- and surgeon-levels changed with the addition of predictors as fixed effects. We conducted likelihood ratio tests, which assesses the difference in two models using the Chi square distribution to evaluate their significance in explaining the total variability in continence outcomes. The significance of specific predictors was assessed using point estimates and 95% confidence intervals. Numeric variables were centered and scaled, and model diagnostic tests were performed to assess if any model assumptions were violated.

3.3.4 Treatment of Missing Data

Prior to constructing the mixed-effects models, we assessed the extent of missing data at multiple time points. As the mixed-effects models effectively handle data missing at random[184], we included all patients who completed at least one PRO survey. However, as complex mixed-effects models often fail to converge, we built the models with a subset of the dataset after removing patients with missing covariates. We conducted sensitivity analyses to assess the robustness of the findings using the full-data set (which included missing data) and a subsequent dataset with imputed data.

3.4 Results

We identified 9,159 men who underwent radical prostatectomy and completed EPIC-26 surveys at multiple time points from 2015 to 2020. Survey completion rates at baseline, 3, 6, 12 and 24 months were 78.2%, 68.3%, 61.3%, 56.2% and 41% respectively. Table 1 summarizes the clinical and demographic characteristics of the study cohort. Briefly, the median age at diagnosis was 64 years (IQR 58, 68). Seventy-six percent of men self-identified as non-Hispanic White and 11% self-identified as Black. Eighty-nine percent of men had 1 or less comorbidity while 10% had 2-3 comorbidities. Median BMI was 28kg/m^2 (IQR 26-32). Similarly, 45% (n=4106) of newly diagnosed men had grade group 2 disease while 22% (n=2006) had grade group 3 disease. Seventy four percent of patients had T1 disease while 12% had T2a and 13% with greater than T2a disease. Sixty eight percent of men undergoing RP had bilateral nerve sparing procedure (Table 1).

Mean urinary continence at baseline was 90 (SD 14.8) on a 0 to 100 scale. Mean urinary continence score decreased to 52 (SD 27.4) at 3 months post-surgery and subsequently increased to 67 (SD 26.1) at 6 months; and 72.3 (SD 24.8) and 73.5 (SD 24.2) at months 12 and 24, respectively. When stratified by surgeon's annualized radical prostatectomy volume, mean urinary domain score at baseline was similar for all 4 strata (Table 2). Mean urinary domain score at 3 months was higher for patients associated with high volume surgeons (52.1, SD 27.4) versus 44.1 (SD 26.8) for low volume surgeons. The absolute difference in scores between high and low volume surgeons persisted across all subsequent time points. On average, patients

operated on by high volume surgeons recovered 83% of their baseline function while patients operated on by low volume surgeons recovered 70% of their baseline function by 24 months after surgery (Table 2).

The analytic cohort created for the mixed-effects model consisted of 6,025 men (Supplementary Figure1). The base model shows the mean urinary continence score with an intercept of 60.95 (CI 59.33 – 62.56) across all time points post-surgery. The overall intraclass correlation coefficient (ICC) in the base model indicates that 57% of variance was attributed to patients and 6% attributed to surgeons (Table 3). The intermediate model shows the mean urinary continence score at month 3 with an intercept of 48.88 (CI 47.26 – 50.49); when time as a categorical variable is added to the model while keeping the random effects structure consistent. Mean urinary continence score increases by 14.95 units to 63.83 at 6 months, further increasing to 69.23 at 12 months and 70.3 at 24 months. In this model, variability attributed to patients and surgeons was 69% and 7% respectively. The final model produced slightly higher estimates at different time points with the mean urinary continence score at month 3 at 56.53 (CI 53.39 – 59.67). The variability attributed to patients declined to 66% while the variability attributed to surgeons remained steady at 7% as we introduced additional predictors in the final model. Additions of age, BMI, Charlson comorbidity index, Grade Group, and clinical T-stage decreased the mean continence scores by -3.13, -1.70, -1.44, -1.81 and -0.41 units, respectively. However, the addition of baseline urinary continence score, the surgeon's annualized radical prostatectomy volume, and whether a bilateral nerve-sparing procedure was performed increased the mean continence by 5.03, 3.18 and 1.22 units, respectively.

3.5 Discussion

We quantified variability attributable to both patients and surgeons using longitudinal patient reported outcomes data. We found that 66% of variability in urinary function recovery was attributable to patients with only 7% of variability attributable to surgeons in the three models. We observed that advancing age, increased BMI, and Charlson comorbidity index were negatively associated with urinary function recovery while increasing time since surgery, surgeon's radical prostatectomy volume, surgical technique (e.g., bilateral nerve-sparing) and

baseline pre-operative continence scores were positively associated with urinary function recovery.

Variability in continence outcome is well documented in urologic literature[14,132,162,165] but the variability attributed to surgeons and patients based on PROs is not widely explored.[164] Cross sectional single surgeon series or high-volume institutional studies have assessed rates of post-surgical incontinence at varying time points – focusing on either early or late recovery but few studies assess the unique recovery trajectory.[172,185] A strength of the present study is the assessment of longitudinal variability in functional recovery across the patient population from diverse practices in a large prostate cancer registry. The present study advances our understanding of variability attributable to patients and surgeons based on routine clinical practice data in several ways. First, few studies incorporate patient reported outcomes to assess continence and urinary function recovery, and of those that use PROs, very few demonstrate the volume-outcome association in post radical prostatectomy continence recovery.[173,174,186] In fact, a recent high-volume center study of 3,945 patients from 2008-2019 showed no measurable improvement in urinary function outcomes post RP, where 80% of cases were performed by experienced surgeons.[163] Second, there is a lack of understanding of what contributes to variability in outcomes. Indeed, it is unknown if variability is attributed to modifiable or unmodifiable patient factors, modifiable surgeon factors or other sources of variability that are not well understood. We have attempted to identify and quantify the sources of variability. Using the same dataset, Auffenberg et al. demonstrated wide variations in reporting of good urinary function at 3 months post-surgery.[168] We built on this finding to assess the variability in urinary function recovery that can be attributed to patients and surgeons given the heterogeneity in long term recovery.

Putting our findings in context against the well-established volume-outcome association is important. Surgeon experience and hospital volume have been shown to reduce adverse outcomes including positive surgical margins, post-operative infections, 30-day readmission, need for adjuvant and salvage therapies, and mortality.[170,175,187,188] Our analyses point to greater variability attributable to patients. Given these findings, one might ask, “why do two patients operated on by the same surgeon using similar surgical techniques have different functional outcomes?” Part of the answer is that there are two sides to assessing outcomes. The technical

surgical side, of which the surgeon has control over, measures how well the surgery was performed. The patient side, of which the surgeon has less control over, measures how well the patient is recovering over time based on patient generated reports. Superior surgical skills are critical to continence recovery[176] but our analyses suggest that patient factors contribute more in this context based on the attributable variability that we observed.

There are several limitations to this study. First, this is an observation study of a state-based disease registry, and the results may not be generalizable to other environments. Second, the present study assesses variability in outcomes based on patient reported outcomes. It is important to recognize, despite being an important endpoint, and EPIC-26 being a validated instrument, some subjectiveness in capturing recovery may be involved. PROs capture patient perspectives and perception of recovery, which are different from biological or clinical factors, and PROs may not be able to capture the entirety of continence and urinary function recovery due to measurement errors or imprecise interpretation of continence. For example, pad weight testing is recognized as a more objective measurement of incontinence, lack of pad testing is an important limitation to this study. Third, factors including membranous urethral length and other anatomic features are known to impact continence recovery post prostatectomy.[110] Lack of anatomical data and other unmeasured confounders that could potentially impact early urinary function recovery is a limitation of our study. Fourth, only surgeon volume and nerve sparing procedures were available to be included in the study to account for surgeon effects. We were not able to assess surgical training and experience. Despite these limitations, this study is one of the first to quantify the variability in urinary function recovery and assess variability explained by patient- and surgeon-level factors.

3.6 Conclusion

The present study broadens our understanding of variability in urinary function recovery associated with patients and surgeons. We observed wide variability in urinary function recovery explained by both patient and surgeon level factors. Variability explained by patient-level factors was greater than surgeon-level factors. Surgeon's annual case volume did not fully explain the variability observed in our models. Efforts to include PROs in the assessment of urinary function recovery is still in its infancy. It is established that higher surgeon volume leads to better

oncologic outcomes. Higher surgeon volume may also lead to better patient functional outcomes though our results did not show a strong association between volume and urinary function recovery. Quantifying variability attributable to patients and surgeons may help to better understand the nature of urinary function recovery from patients' perspective. It may also lead to quality improvement opportunities and support treatment decision-making process by setting realistic expectations of early urinary function recovery while counseling candidates for radical prostatectomy.

Table 1. Clinical and Demographic Characteristics of Study Sample

Characteristic	Median (IQR) or No. (%)
Age, median (IQR), y	64 (58 – 68)
BMI, median (IQR)	28 (26 – 32)
Charlson Comorbidity Index, N(%)	
0-1	8190 (89.4)
2-3	899 (9.8)
3+	66 (0.7)
Race	
Black	1002 (10.9)
White	6991 (76.3)
Other	210 (2.3)
Unknown	956 (10.5)
Gleason Grade Group	
1	1461 (16.0)
2	4106 (44.8)
3	2006 (21.9)
4	876 (9.6)
5	627 (6.8)
Missing	83 (0.9)
Clinical stage	
T1	6849 (74.8)
T2a	1095 (12.0)
>T2a	1189 (13.0)
Missing	26 (0.3)
Nerve Sparing	
No	2968 (32.4)
Bilateral	6191 (67.6)

Table 2. Urinary Domain Scores Stratified by Surgeon Volume

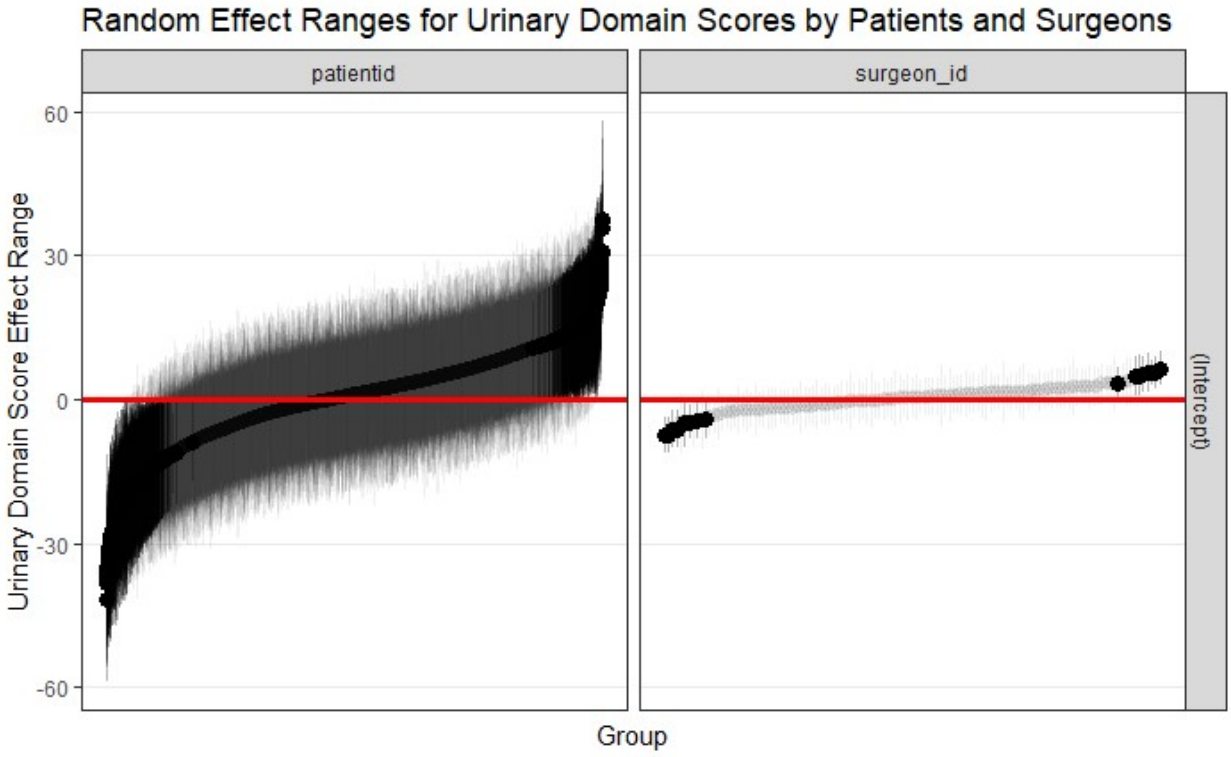
	Low Volume Quartile 1 (n=32)	Mid Low Volume Quartile 2 (n=32)	Mid Volume Quartile 3 (n=31)	High Volume Quartile 4 (n=32)	Overall (N=127)
Total Patients (N)	368	789	2007	5990	9159
Baseline					
Mean (SD)	89.0 (16.2)	88.2 (17.0)	89.8 (14.2)	89.8 (14.6)	89.6 (14.8)
Missing	106 (28.8%)	183 (23.2%)	512 (25.5%)	1196 (20.0%)	1999 (21.8%)
Month 3					
Mean (SD)	44.1 (26.8)	46.9 (26.5)	51.7 (27.4)	53.2 (27.4)	52.1 (27.4)
Missing	163 (44.3%)	285 (36.1%)	684 (34.1%)	1771 (29.6%)	2904 (31.7%)
Month 6					
Mean (SD)	58.9 (26.9)	62.3 (27.4)	65.9 (26.1)	67.7 (25.7)	66.6 (26.1)
Missing	180 (48.9%)	334 (42.3%)	809 (40.3%)	2222 (37.1%)	3547 (38.7%)
Month 12					
Mean (SD)	64.7 (27.2)	68.6 (26.3)	71.8 (24.8)	73.1 (24.4)	72.3 (24.8)
Missing	218 (59.2%)	397 (50.3%)	920 (45.8%)	2474 (41.3%)	4010 (43.8%)
Month 24					
Mean (SD)	62.9 (28.2)	71.2 (25.5)	71.9 (24.4)	74.7 (23.7)	73.5 (24.2)
Missing	250 (67.9%)	534 (67.7%)	1221 (60.8%)	3391 (56.6%)	5400 (59.0%)

Table 3. Variability in Urinary Domain Scores After Radical Prostatectomy in MUSIC

	Base Model		Intermediate Model		Full Model	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*	*Estimates*	*CI*
Fixed effects						
Intercept	60.95	59.33 – 62.56	48.88	47.26 – 50.49	56.53	53.39 – 59.67
Mo. 6 Urinary Function			14.95	14.41 – 15.49	15.04	14.43 – 15.50
Mo. 12 Urinary Function			20.35	19.79 – 20.91	20.37	19.77 – 20.91
Mo. 24 Urinary Function			21.32	20.70 – 21.93	21.34	20.60 – 22.04
Age at Prostatectomy					-3.13	-3.60 – -2.43
Body Mass Index					-1.70	-2.30 – -1.17
Charlson Comorbidity Index					-1.44	-2.09 – -0.74
Baseline Urinary Function					5.03	4.62 – 5.79
Surgeon Volume					3.18	0.78 – 5.62
Nerve Sparing (Bilateral vs. None)					1.22	-0.32 – 2.50
Gleason Grade Group					-1.81	-2.31– -1.19
Prostate Specific Antigen					-0.29	-0.88 – 0.31
Clinical T-stage					-0.41	-0.89 – 0.08
Random effects						
Residual σ^2	268.59		164.80		164.85	
Patient Intercept τ_{00}	419.31		453.17		403.03	
Surgeon Intercept τ_{10}	46.68		45.05		41.02	
ICC (Patients) τ_{01}	0.57		0.69		0.66	
ICC (Surgeons) τ_{11}	0.06		0.07		0.07	
Patients (n)	6025		6025		6025	
Surgeons (n)	118		118		118	
Observations	16955		16955		16955	
Marginal R^2	0.00		0.105		0.18	
Conditional R^2	0.63		0.78		0.78	

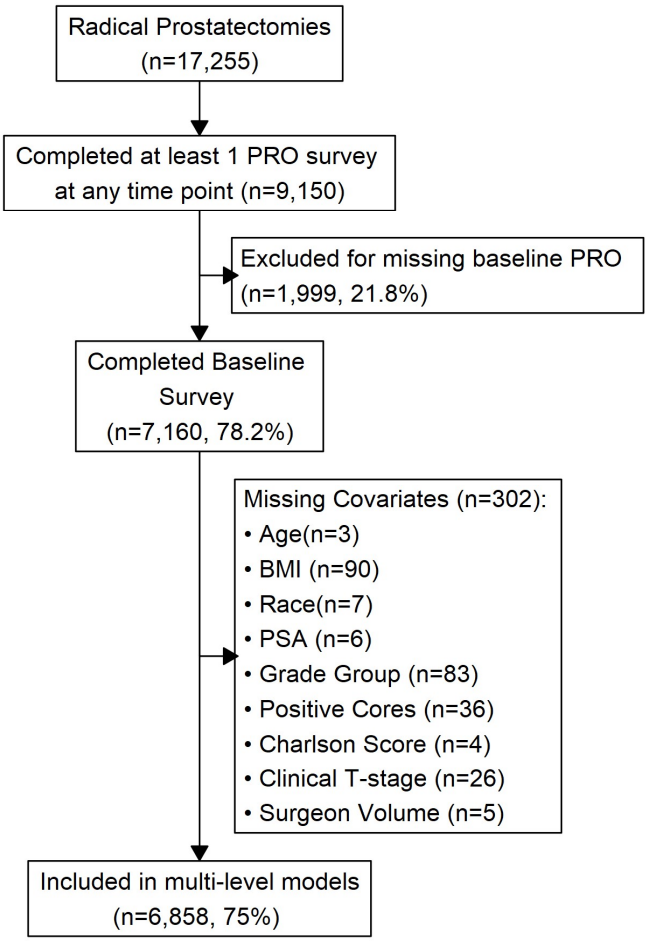
ICC: Intraclass correlation coefficient; R2: Coefficient of determination ; Marginal R2: Variance explained by only the fixed-effects structure
Conditional R2: Variance explained by both the fixed and random effects structures

Figure 3. Effect ranges based on model outputs from the full model.



Panel on left shows deviations from the overall mean urinary domain scores for patients in MUSIC registry. Panel on right shows deviations from the overall mean urinary domain scores for surgeons in MUSIC registry.

Figure 4. Consort Diagram of the Study Cohort



Chapter 4 Dynamically Predict the Long-term Recovery of Urinary Continence Following Radical Prostatectomy Using Preoperative and Postoperative Data

4.1 Abstract

Objective: To predict long-term recovery of urinary continence after radical prostatectomy using clinical and patient reported outcomes data.

Methods: Using pre- and postoperative clinical and patient reported outcomes data from the Michigan Urological Surgery Improvement Collaborative (MUSIC), we trained and validated dynamic prediction models to estimate the probability of achieving continence at multiple postoperative time points. We trained two random forest models that predicted urinary continence outcomes at 3-, 6-, 12- and 24-months after radical prostatectomy. Primary outcomes were patient reported urinary domain scores and pad use. Pad use was dichotomized as 0 pad (continent) per day vs 1 or more pads (incontinent) per day. Urinary domain scores > =74 was considered as good urinary function. Men who underwent radical prostatectomies between September 2016 and October 2021 and completed patient reported outcomes survey were included in the study.

Results: Preoperative model with clinical and baseline urinary function information performed poorly predicting both urinary domain scores and pad use at 3 months with an AUC of 0.75 and 0.59, respectively. Preoperative model predicted both outcomes at 6 months with an AUC of 0.70 and 0.60. As the prediction window increased, predictive performance declined for the preoperative model. In contrast, the postoperative model performed better with the addition of patient reported outcomes data. The postoperative 3-month model predicting outcomes at 6 months and the 6-month model predicting urinary domain scores at 12 months performed better with an AUC of 0.86 and 0.88, respectively. Similarly, the 3-month model predicting pad use at 6-moth performed well with an AUC of 0.92 and the 6-month model predicting pad use at 12-month performed modestly better with an AUC of 0.93.

Conclusion: Setting realistic expectations of urinary function recovery is an important aspect of patient counseling and shared decision-making. Updating expectations throughout the recovery period is a challenge as prediction tools in prostate cancer care have traditionally used pre-operatively available information to predict postoperative outcomes. Integrating patient reported outcomes to predict long-term recovery of urinary continence after radical prostatectomy improves predictions, thereby supporting patient counseling and setting realistic expectations of urinary continence recovery.

4.2 Introduction

Men considering radical prostatectomy (RP) after a prostate cancer diagnosis face a spectrum of potential functional and oncological outcomes. Functional recovery after RP varies based on factors including a patient's anatomy, tumor biology, clinical and demographic characteristics, as well as the surgeon's skill and surgical technique.[1–4] Risk prediction models estimate functional recovery after radical prostatectomy,[5–7] but are typically designed to support treatment planning using preoperative data, thereby missing an opportunity to incorporate post-treatment quality of life information to support personalized survivorship care.[7–9]

One of the most difficult challenges is setting realistic expectations of functional recovery and updating those expectations throughout the recovery period. Models that use preoperative variables to estimate post-operative outcomes often do not perform well.[11] For patients who do not experience early recovery of continence after radical prostatectomy, the focus shifts to what degree of recovery remains possible. Thus, preoperative predictions should be updated during the postoperative period as new information becomes available. By adding new information, the postoperative models may better characterize the trajectory of patients' recovery based on the extent of recovery that has already occurred.[6,11]

In this chapter, we develop and validate prediction models to estimate the probability of achieving continence at 3-, 6-, 12-, and 24-months after RP by incorporating longitudinal data to predict outcomes at multiple time points.[6,10–12] The models allow for the use of either pre- or post-operative demographic, clinical, tumor, and patient reported outcomes (PRO) data. We hypothesize that predictions made post-operatively will be increasingly accurate as compared to preoperative predictions. In patients who recover faster or slower than initially expected, post-

operative predictions may be used to reassure patients, to reset their expectations, or to recommend rehabilitative or surgical interventions.

4.3 Methods

4.3.1 Data Sources

The Michigan Urological Surgery Improvement Collaborative (MUSIC) is a physician-led urologic quality improvement collaborative that maintains a prospectively updated registry that collects clinical, demographic and patient reported outcomes data from men diagnosed with prostate cancer. Trained abstractors review electronic health records at their respective urology practices and periodically submit diagnosis-related data to MUSIC. Men diagnosed with prostate cancer are encouraged to complete the Expanded Prostate Cancer Index Composite-26 (EPIC-26) PRO survey preoperatively and post-operatively at 3-, 6-, 12- and 24-months. We used both the abstracted clinical data from the MUSIC registry and the EPIC-26 data that is part of the MUSIC-PRO program. All newly diagnosed prostate cancer patients who underwent RP and completed EPIC-26 surveys at baseline, 3-, 6-, 12-, and 24-months post RP were included in the study. Participating MUSIC practices obtained approval or exemption from their respective institutional review boards to participate in a statewide quality improvement collaborative.

4.3.2 Study Cohort

Derivation and Validation Cohorts:

We included men in the MUSIC registry who were diagnosed with prostate cancer and had radical prostatectomy performed between September 2016 and October 2021. We randomly split newly diagnosed men who completed EPIC-26 surveys at baseline and at any post-RP time points into derivation (66%) and validation (33%) cohorts. The split was stratified by surgeons with each surgeon appearing in both cohorts; however, each patient was only included in either cohort. Men who did not complete any postoperative EPIC-26 surveys (either 3-, 6-, 12- or 24-months) but completed only the baseline survey were excluded from the study.

4.3.3 Outcomes

Our primary outcomes of interest were urinary continence at four post-operative timepoints (i.e., 3-, 6-, 12-, and 24-months), defined in two ways: (a) Urinary incontinence domain scores calculated as a composite score ranging from 0-100 in EPIC-26, with higher numbers reflecting better self-reported urinary function and a score of >=74 clinically considered as good urinary function,[13,14] and (b) Pad use; based on patient responses to the EPIC-26 item that asks for the number of pads or adult diapers used to control leakage during the past 4 weeks. Responses were dichotomized into 0 pads per day (continent) versus 1 or more pads per day (incontinent). We assessed the performance of the models to predict urinary domain scores and pad use outcomes at each time point after radical prostatectomy.

4.3.4 Predictors

Time-invariant predictors, or those that did not change over time, included age at radical prostatectomy, BMI, clinical T-stage, preoperative PSA, Gleason grade group, nerve sparing status and prostate gland volume. Time-varying predictors, or those that changed over time, included EPIC-26 urinary domain score and pad use status, which were collected preoperatively and at time points 3-, 6-, and 12-months after surgery.

4.3.5 Model Development

In the absence of missing data, each patient contributed up to four observations to the model development process: preoperative data to predict 3-month outcomes, 3-month data to predict 6-month outcomes, 6-month data to predict 12-month outcomes, and 12-month data to predict 24-month outcomes. Using the time-varying and time-invariant variables, we trained models to predict urinary continence recovery at four time points. Because a single model is used to generate predictions pre- and post-operatively at multiple time points, we refer to this as a dynamic model.

We built two dynamic random forest (RF) prediction models; one to predict urinary domain score (continuous) and the other to predict pad use after surgery (dichotomous). The Random forests, initially proposed by Brieman[189], are ensembles of decision trees that are trained following a 2-step process. The first step is a process called bagging, which builds a set of decision trees on bootstrapped samples and aggregates the results from the different samples.[190]

The second step involves implementing randomness of feature selection, which ensures that no one predictor has an outsized dominance on the classification by forcing each split to consider only a subset of predictors and can only use one of those predictors at each split.[191],[192] The RF models extend many benefits. First, RF models are flexible and can handle both classification and regression problems. Variable importance is easier to assess in RF models. Since they use a random subset of decision trees to generate predictions, they reduce but not completely eliminate the risk of over-fitting. Despite these benefits, RF models have some key drawbacks. In noisy datasets, RF models could split nodes along less meaningful data and could produce inaccurate or overfitted predictions. RF models may also be slow to run on large data sets and are often difficult to interpret.

We trained the two RF models (i.e., one for each outcome) on the derivation set and evaluated the performance on the validation set. The first RF model was trained to predict EPIC-26 urinary continence domain scores at 3-, 6-, 12-, and 24-months post RP. Each model included both time-invariant data and time-varying urinary function outcomes data. Using the same modeling strategy, the second random forest model was built to predict pad use at 3-, 6-, 12-, and 24-months post RP.

4.3.6 Model Performance

We evaluated the performance of our models in the validation cohort by assessing model discrimination and calibration.

Urinary Domain Score Prediction: EPIC-26 urinary continence domain score was evaluated based on the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) scores and increasing AUC scores for urinary domain scores. We considered good urinary function to mean that EPIC-26 urinary domain scores were 74 and higher based on previous work.[13,16] We used calibration plots to compare between predicted values against observed values.[13]

Pad Use Prediction: We used the area under the receiver operating characteristic curve (AUC) to assess the discrimination of our model predicting pad use status. We assessed the AUCs at 3-,6-, 12-, and 24-months post RP. We considered men to be continent if they reported using 0 pads per day in the past 4 weeks based on definitions. This is a more restrictive definition

than prior studies. We opted for a stricter definition because studies have shown considerable decrease in quality of life between 0 pads versus one or less pads per day definition.[193]

4.3.7 Missing Data

In the *h2o* implementation of random forests, missing values are assumed to carry valuable information.[20] The algorithm treats missing values as a separate category that can go either left or right during tree building stages, similar to how any other categorical values are handled during the splitting process.

4.3.8 Software

All analyses were conducted in R 4.1.2 (R Foundation, Vienna, Austria). Data were prepared using the *tidymodels* package and models were fit directly in *h2o*, an open-source, machine learning and predictive analytics platform using the *h2o* R package. Model performance was visualized using the *runway* package.

4.4 Results

4.4.1 Population Characteristics

We identified 4,943 men who completed at least one EPIC-26 survey with at least 4 responses to the urinary domain questions, which was required for the calculation of urinary domain scores. Of these, 3,295 were randomly assigned to the derivation cohort and 1,648 were assigned to the validation cohort. The derivation cohort included 3,280 (3-month), 2,519 (6-month), 2,199 (12-month) and 1,657 (24-month) men, where the time refers to the number of months post-RP. The validation cohort included 1,638 (3-month), 1,267 (6-month), 1,114 (12-month) and 860 (24-month) men. Detailed characteristics of men who completed the EPIC-26 surveys are presented in Table 1. Characteristics of men stratified by pad use outcome at months 3 and 12 are presented in Tables 2a and 2b, respectively. Briefly, men who remained incontinent were older, with a median age of 66 (IQR 61,70, $p < 0.001$) as compared to those who were continent, who had a median age of 63 (58,67). Incontinent men had higher gland volume of 38

mL (IQR 29, 50) vs 35 mL (IQR 27, 47, p <0.001). Greater proportion of men who were incontinent at 3 month had incomplete or no nerve sparing 43% vs 36% (p <0.001). Additionally, men who remained incontinent at month 3 had greater proportions of grade group 3, 4 and 5 disease (Table 2a). These differences in grade group diseases between incontinent and continent men were also seen at 12 month (Table 2b).

4.4.2 Prevalence of Outcomes

EPIC-26 urinary domain score. At baseline, the median EPIC-26 urinary domain score was 100 (IQR 86,100). Post-operative median domain scores ranged from 52 (IQR 31,71) at 3 months; 67 (IQR 46,92) at 6 months; 79 (IQR 54,100) at 12 months; and 79 (IQR 59,100) at 24 months. Similarly, 85.5% (n=3,444) of men had a good urinary domain score defined by a score of ≥74 at baseline, 20% of men had good urinary function at 3 months, which subsequently increased to 41% at 6 months, 53% at 12 months and 54% at 24 months.

Pad use. Ninety-eight percent (n=3,954) of men were pad-free at baseline in the overall cohort. At 3 months, 73% (n=3,577) of men required one or more pads per day (incontinent). Pad use decreased to 50% (n=1,888) at month 6, with further decline 37% (n=1,210) at month 12; and plateaued at 31% (n=789) through month 24 in the combined derivation and validation cohorts.

4.4.3 Model Performance in the Validation Cohort

EPIC-26 urinary domain scores.

Baseline Model: The baseline model with preoperative variables predicted urinary domain score at 3 months with a root mean squared error (RMSE) of 29.13, mean absolute error (MAE) of 23.87, and the area under the curve (AUC) of 0.62 (Table 3). RMSE and MAE *decreased* while AUC *increased* to 0.64 for the baseline model predicting urinary domain scores for 6-months. Slight increases in RMSE and MAE values were observed for the baseline model predicting outcomes at more distal time points of 12- and 24-months, while AUC decreased to 0.62 reflecting poor performance of the baseline model (Table 3).

Post-operative model: Models that incorporated post-operative data and predicted outcomes at proximal time points [i.e., month 3 model predicting outcomes at month 6 (RMSE 16.60; MAE 13.22; AUC 0.88) and month 6 model predicting outcomes at month 12 (RMSE; 16.67 and MAE; 12.69; AUC 0.86)] performed better than baseline models or models that predicted outcomes at distal time points. The 12 month model predicting 24 urinary domain scores performed the best (RMSE of 16.01; MAE 12.24; AUC 0.87).

Pad Use

Our baseline model predicted pad use with an AUC of 0.62 at 3-months; 0.63 at 6 months; 0.60 at 12 months; and 0.63 at 24 months post RP. Prediction of pad use also improved with the addition of post-operative data, including pad use at previous time points. For example, the 3-month model predicting pad use at 6-month performed well with an AUC of 0.89. Similarly, the 12-month model predicting pad use at 24-month performed well with an AUC of 0.92 while the baseline model predicting 12-month outcomes performed the worst with an AUC of 0.60. The post-operative models performed slightly better for pad use prediction at more proximal outcomes except for the baseline models.

We assessed model calibration by plotting predicted versus observed EPIC-26 urinary domain scores and predicted vs observed pad-use at 3-, 6-, 12- and 24-months. Calibration plots show the model predicting urinary domain score overestimates risk at month 3 and under-estimates at 6, 12 and 24 months (Figure 1). The model predicting pad use also over-estimates risk at month 3 and underestimates at month 6, 12, and 24.

4.4.4 Sensitivity Analyses

We conducted sensitivity analyses for both outcomes by adding surgeon's annualized surgical volume and whether bilateral nerve sparing was performed as predictors. Addition of surgical volume did not improve model performance at months 3, 6, 12 or 24. The RMSE and MAE values were 17.90 and 14.85 respectively for the baseline model predicting 3-month outcome and with an AUC of 0.75 for good urinary function. For the 3-month model predicting 6 month outcomes, the RMSE and MAE were 14.19 and 11.28 respectively with an AUC of 0.89 for good urinary function, virtually no change in model performance from the original model (Table 8). Addition of both nerve sparing and surgeon volume did not improve model

performance. RMSE, MAE and AUC values remained unchanged for baseline and post-operative models indicating the addition of nerve sparing information did not improve model performance. When all predictors including nerve sparing and surgeon volume were added, model performance improved slightly (AUC 0.76 vs. 0.75 in the original model) for baseline model predicting outcomes at 3 and 6 months (AUC 0.71 vs 0.70 in the original model) respectively.

4.5 Discussion

In this study, we trained dynamic models to predict urinary domain scores and pad use among patients undergoing RP. Our models incorporating post-operative data in addition to the preoperatively available data performed better than models trained only on preoperative data; model performance improved with the addition of new data. We also found that post-operative models that predict short-term outcomes provide more precise estimates of urinary function recovery as compared to long-term outcomes. In comparison to the post-operative models, the preoperative baseline models performed poorly in all instances. Long-term pad use prediction improved with the availability of post-operative data, with the best estimates of 24-month pad use arising from models that incorporate predictors from 12 months.

Our results are consistent with previous findings that have used post-operative data. Beyond continence recovery, dynamic models incorporating pre- and post-treatment data have shown to improve performance of models predicting sexual function recovery and oncologic outcomes.[22,24,25] Finelli et al. found that dynamic prediction models had better performance predicting radiographic progression free survival (PFS) when using longitudinal PSA data in metastatic castration resistant prostate cancer setting.[22] Similarly, Agochukwu-Mmonu et al. found that the inclusion of post RP data improved the performance of multivariate models predicting sexual function recovery in men undergoing radical prostatectomy in the MUSIC registry.[24] Likewise, using pre- and post-operative models predicting continence after radical prostatectomy, Tutolo et al., found that models that incorporated 3-month post-operative data to the baseline data showed better performance than model that relied only on baseline preoperative predictors.[23] Lastly, Vickers et al. also found patient-reported functional status in the first post-operative year to be highly predictive of future functional status.[194]

Evidence underscores the importance of using both pre- and post-operative data to improve accuracy for more personalized prediction.[11,15,21–23] Further, post-operative models establish realistic expectations of long term functional recovery. Realistic expectations of urinary function and pad use pre- and post-operatively support treatment decision-making process and reduces the potential for decision regrets as poor urinary function is one of the most bothersome early adverse events for men undergoing RP.[195,196] Ability to predict and subsequently intervene to restore continence is an important aspect of survivorship care. However, studies have shown that simple immediate post-operative functional outcome recovery estimates are not meaningful to those who have yet to gain continence post operatively.[26] Continence recovery is a long-term process with continual gains even after the 12 month period when most men are expected to regain baseline continence levels.[27] Hence, it is important to build predictive tools that provide estimates of continence over a longer time horizon incorporating both posttreatment data and postoperative time to provide more accurate estimates.[15,26]

Our study has some limitations. First, anatomical features and surgical skills have been considered to impact early continence recovery despite variability in outcomes by among surgeons.[4,28–30] We are limited by the lack of anatomical data to incorporate in our prediction models. However, we did not find improvement of model performance when adding surgical volume as a predictor. Second, we did not include pathological features in our modeling, which could potentially further improve our prediction accuracy. However, use of pathological data may not be readily available as input data during patient counseling. Despite these limitations, the present study advances our understanding of the potential role of dynamic models in prostate cancer care. As most contemporary prediction models are built for preoperative counseling, the present study adds to the growing literature supporting the use of the dynamic modeling approach.[24] Dynamic models are not only useful for treatment making process to illustrate potential recovery trajectories during counseling but also as a post-operative tool to assess if early interventions to improve continence are appropriate.

4.6 Conclusion

Setting realistic expectations of functional recovery is one of the most important challenges for clinicians. While we found preoperative data to be unreliable in predicting long-

term continence, incorporation of post-operative data using dynamic modeling can help identify men who might benefit from early post-operative interventions to improve urinary functions after radical prostatectomy.

Table 4. Patient characteristics at baseline stratified by cohorts

	Derivation Cohort		Validation Cohort	
Variable	N	N = 3,295[1]	N	N = 1,648[1]
Age	3,295	65 (60, 69)	1,648	65 (60, 69)
BMI	2,645		1,299	
<25		388 (15%)		208 (16%)
25-30		1,206 (46%)		558 (43%)
>30		1,051 (40%)		533 (41%)
Unknown		650		349
Race	3,295		1,647	
White		2,593 (79%)		1,260 (77%)
Black		253 (7.7%)		149 (9.0%)
Other		51 (1.5%)		26 (1.6%)
Unknown		398 (12%)		213 (13%)
Gland Volume	2,239	37 (28, 50)	1,102	36 (28, 50)
Unknown		1,056		546
PSA	2,233		1,112	
< 4		315 (14%)		148 (13%)
4-10		1,435 (64%)		722 (65%)
>10		483 (22%)		242 (22%)
Unknown		1,062		536
Clinical T-Stage	3,082		1,525	
T1		2,354 (76%)		1,153 (76%)
T2a		328 (11%)		188 (12%)
>T2a		400 (13%)		184 (12%)
Unknown		213		123
Nerve Sparing	3,295		1,648	
Incomplete		1,349 (41%)		691 (42%)
Complete (Bilateral)		1,946 (59%)		957 (58%)
Grade Group	2,853		1,422	
1		395 (14%)		202 (14%)
2		1,289 (45%)		656 (46%)
3		672 (24%)		324 (23%)
4		295 (10%)		150 (11%)
5		202 (7.1%)		90 (6.3%)
Unknown		442		226

[1]Median (IQR) or Frequency (%)

Table 5. Characteristics of study participants by pad use outcome at month 3

Characteristics	N	Overall, N = 4,915[1]	Month 3 Pad Use: Continence, N = 1,337[1,3]	Month 3 Pad Use: Incontinence, N = 3,586[1]	p-value[2]
Age	4,915	65 (60, 69)	63 (58, 67)	66 (61, 70)	<0.001
BMI	3,925				0.005
<25		593 (15%)	145 (14%)	448 (15%)	
25-30		1,754 (45%)	506 (49%)	1,284 (41%)	
>30		1,578 (40%)	382 (37%)	1,196 (41%)	
Unknown		990	304	686	
Race	4,914				0.0087
White		3,829 (78%)	1,052 (79%)	2,777 (78%)	
Black		400 (8.1%)	91 (6.8%)	309 (8.6%)	
Other		77 (1.6%)	27 (2.0%)	50 (1.4%)	
Unknown		607 (12%)	167 (12%)	441 (12%)	
Gland Volume	3,321	37 (28, 50)	35 (27, 47)	38 (29, 50)	<0.001
Unknown		1,594	485	1109	
PSA	3,330				0.26
<4		463 (14%)	144 (15%)	319 (13%)	
4-10		2,147 (64%)	601 (64%)	1,546(65%)	
>10		720 (22%)	193 (21%)	527 (22%)	
Unknown		1,585	399	1,186	
Clinical T	4,581				0.011
T1		3,490 (76%)	987 (79%)	2,503 (75%)	
T2a		508 (11%)	121 (9.7%)	387 (12%)	
>T2a		583 (13%)	137 (11%)	446 (13%)	
Unknown		334	92	242	
Nerve Sparing	4,915				<0.001
Incomplete		2,021 (41%)	478 (36%)	1,543 (43%)	
Complete (Bilateral)		2,894 (59%)	859 (64%)	2,035 (57%)	
Grade Group	4,251				<0.001
1		594 (14%)	198 (17%)	396 (13%)	
2		1,935 (46%)	552 (48%)	1,383 (45%)	
3		991 (23%)	243 (21%)	748 (24%)	
4		440 (10%)	111 (9.6%)	329 (11%)	
5		291 (6.8%)	51 (4.4%)	240 (7.8%)	
Unknown		664	182	482	

[1]Median (IQR) or Frequency (%)

[2]Wilcoxon rank sum test; Pearson's Chi-squared test

[3] Continence is defined as 0 pad per 24 hour period

Table 6. Characteristics of study participants stratified by pad use outcome at month 12

Variable	N	Overall, N = 3,310[1]	Month 12 Pad Use		p-value[2]
			Continence, N = 2,100[1,3]	Incontinence, N = 1,210[1]	
Age	3,310	65 (60, 69)	64 (59, 69)	66 (62, 70)	<0.001
BMI	2,640				0.002
<25		400 (15%)	264 (16%)	136 (14%)	
25-30		1,184 (45%)	783 (47%)	401 (42%)	
>30		1,056 (40%)	627 (37%)	429 (44%)	
Unknown		670	426	244	
Race	3,309				0.012
White		2,644 (80%)	1,688 (80%)	956 (79%)	
Black		222 (6.7%)	119 (5.7%)	103 (8.5%)	
Other		45 (1.4%)	28 (1.3%)	17 (1.4%)	
Unknown		398 (12%)	264 (13%)	134 (11%)	
Gland Volume	2,189	37 (28, 50)	37 (28, 48)	38 (29, 51)	0.045
Unknown		1,121	744	377	
PSA	2,307				0.057
<4		321 (14%)	203 (14%)	118 (14%)	
4-10		1,484 (64%)	974 (66%)	510 (61%)	
>10		502 (22%)	300 (20%)	202 (24%)	
Unknown		1,003	623	380	
Clinical T Stage	3,082				0.014
T1		2,345 (76%)	1,526 (78%)	819 (73%)	
T2a		352 (11%)	206 (10%)	146 (13%)	
>T2a		385 (12%)	230 (12%)	155 (14%)	
Unknown		228	138	90	
Nerve Sparing	3,310				<0.001
Incomplete		1,333 (40%)	795 (38%)	538 (44%)	
Complete (Bilateral)		1,977 (60%)	1,305 (62%)	672 (56%)	
Grade Group	2,890				<0.001
1		404 (14%)	282 (16%)	122 (11%)	
2		1,306 (45%)	850 (47%)	456 (42%)	
3		655 (23%)	394 (22%)	261 (24%)	
4		310 (11%)	182 (10%)	128 (12%)	
5		215 (7.4%)	108 (5.9%)	107 (10%)	
Unknown		420	284	136	

[1]Median (IQR) or Frequency (%)
[2]Wilcoxon rank sum test; Pearson's Chi-squared test; Fisher's exact test
[3] Continence is defined as 0 pad per 24 hour period

Table 7. Performances of models predicting urinary domain scores and pad use by time

Time *(in Months)*		Urinary Domain Scores			Pad Use*
Prediction Time	Outcome Time	RMSE	MAE	AUC Binary	AUC Binary
0**	3	29.13	23.87	0.62	0.62
0	6	25.25	20.80	0.64	0.63
0	12	26.12	22.14	0.62	0.60
0	24	26.52	22.44	0.62	0.63
3	6	16.60	13.22	0.88	0.89
3	12	17.75	13.93	0.84	0.85
3	24	19.02	14.79	0.84	0.85
6	12	16.67	12.69	0.86	0.90
6	24	17.64	13.32	0.85	0.89
12	24	16.01	12.24	0.87	0.92

* Continent = 0 pad use
** 0: Baseline

Table 8. Sensitivity Analysis - Surgeon Volume Included

		Urinary Domain Scores			Pad Use*
Prediction Time	Outcome Time	RMSE	MAE	AUC Binary	AUC Binary
0**	3	17.90	14.85	0.75	0.60
0	6	18.65	15.96	0.71	0.63
0	12	19.14	16.16	0.69	0.60
0	24	19.05	16.14	0.69	0.63
3	6	14.19	11.28	0.89	0.92
3	12	15.59	12.48	0.85	0.91
3	24	16.06	12.86	0.84	0.88
6	12	14.57	11.37	0.88	0.93
6	24	15.08	11.86	0.86	0.89
12	24	14.14	11.14	0.88	0.91

Table 9. Sensitivity Analysis - Nerve Sparing Included

Time *(in Months)*		Urinary Domain Scores			Pad Use*
Prediction Time	Outcome Time	RMSE	MAE	AUC Binary	AUC Binary
0**	3	17.91	14.85	0.75	0.60
0	6	18.65	15.96	0.71	0.64
0	12	19.11	16.13	0.69	0.60
0	24	19.03	16.11	0.69	0.63
3	6	14.11	11.24	0.89	0.92
3	12	15.58	12.46	0.85	0.91
3	24	16.07	12.88	0.85	0.88
6	12	14.55	11.37	0.88	0.93
6	24	15.06	11.88	0.86	0.89
12	24	14.20	11.18	0.88	0.91

Figure 5 Calibration plot for urinary domain score prediction for each time point.

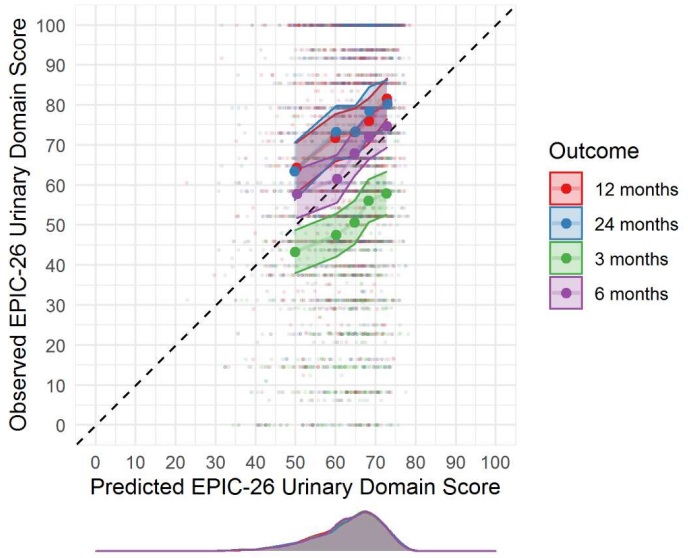


Figure 6. Calibration plot of pad use prediction at each time point

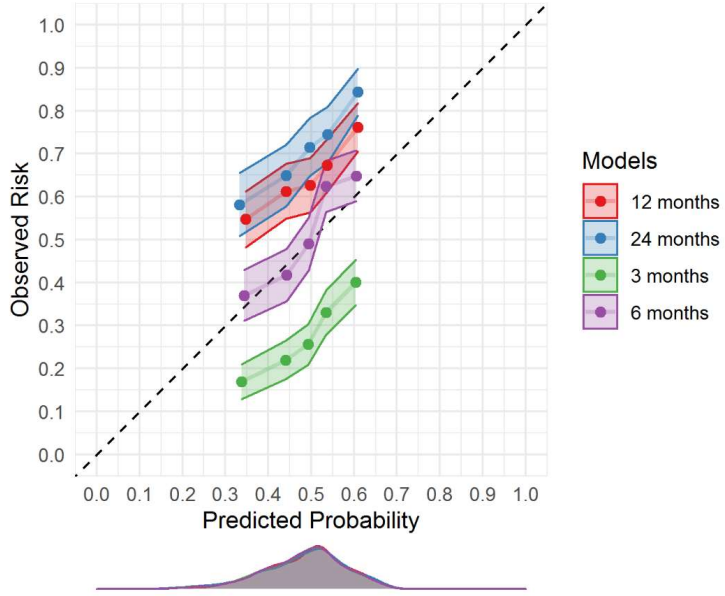


Figure 7. Development of Analytic Cohort

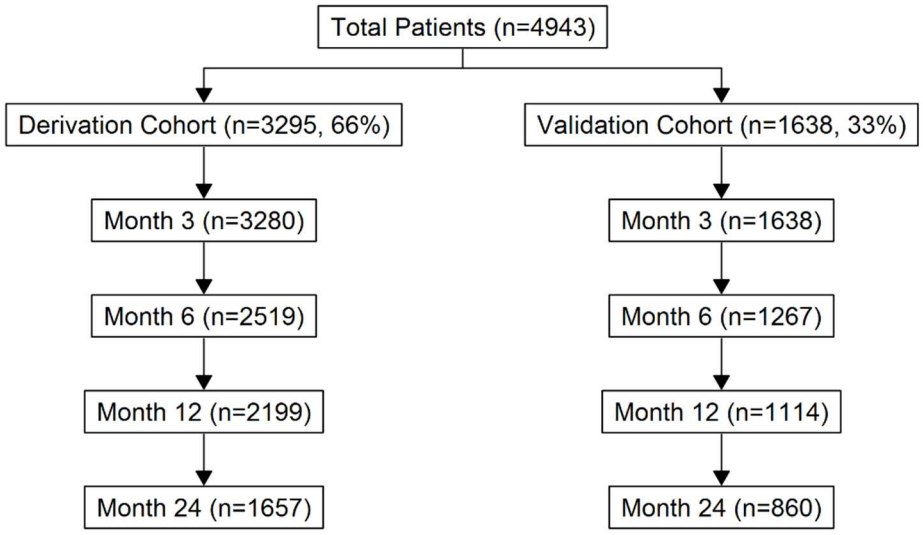
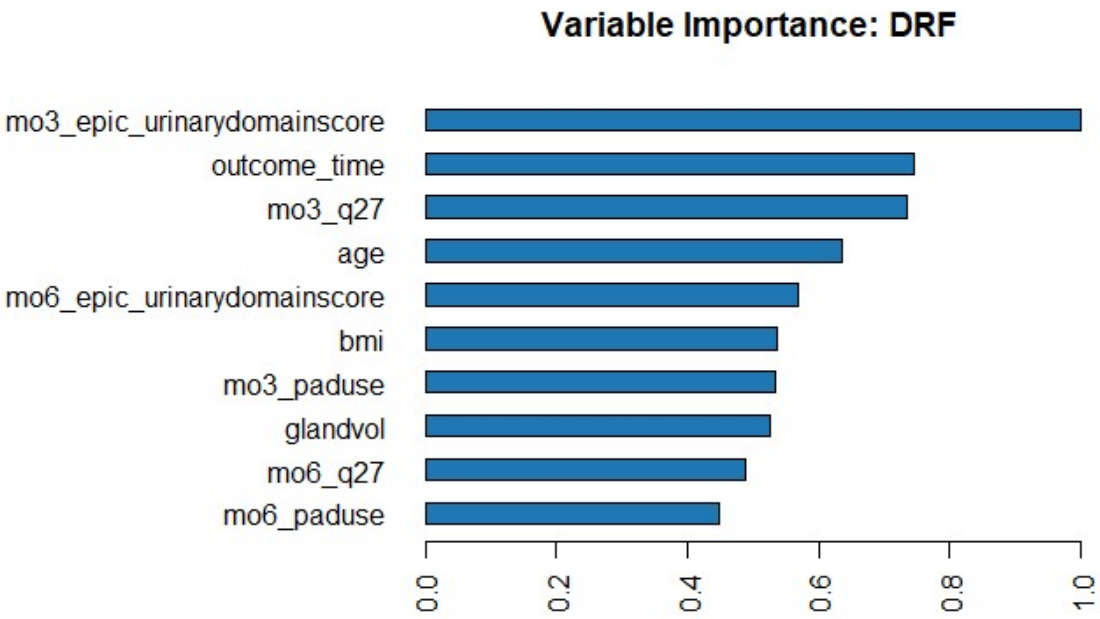


Figure 8. Variable Importance for Pad Use Prediction Model



Chapter 5 Assessing How Secular Changes in Practice Patterns Affect the Performance of Prediction Models Developed from Disease Registry Data

5.1 Abstract

Objective: To assess how temporal changes in practice patterns affect the performance of prediction models trained on disease registry data.

Methods: We used the Michigan Urological Surgery Improvement Collaborative (MUSIC) registry to assess temporal changes in practice patterns, trained logistic regression models using the MUSIC Patient Reported Outcomes (PRO) data to predict the probability of urinary incontinence at 3 months after radical prostatectomy (RP). We included men who underwent radical prostatectomies between January 2017 and December 2021 and completed the MUSIC PRO questionnaire at baseline and 3 months after RP. We assessed model discrimination and calibration over multiple years by comparing baseline approach (no model retraining) against two model updating strategies (models re-trained every year using prior year data and models retrained on aggregate data).

Results: The rates of radical prostatectomy steadily declined in MUSIC from 43% in 2016 to 32% in 2021 with an annual percent change of -5.66. The rates of low-risk disease decreased while intermediate risk disease increased over time. Model discrimination was modest and stable across 3 model retraining approaches. The area under the receiver operating curve (AUC) ranged from 0.59 to 0.63 with no statistically significant differences between models' discriminative performance. Model calibration assessed with intercept and slope remained stable with some indication of over-estimation in later years, but no evidence of calibration drift emerged. Visual inspection of calibration plots also revealed no clear patterns of calibration drift in segments of the curves with most of the predicted probabilities.

Conclusion: We found no indication of calibration drift in prediction models trained on registry data. Models trained on registry data may be less prone to deterioration in model

performance. Future studies should confirm our findings with a larger sample size and assess model deterioration over a longer study period.

5.2 Introduction

Clinical prediction models may deteriorate when the conditions present during model development change after the models are deployed.[123,197] In recent years, researchers have highlighted the phenomenon of dataset shift, which results in deterioration in model performance over time. Dataset shift can occur when there are changes to the distribution of underlying data.[197] After prediction models have learned the relationships between predictors and outcomes in a dataset, changes to the distribution of that data occurring after deployment can result in changes in model performance.[144,197] Dataset shift can occur abruptly, rapidly, gradually, or periodically (i.e., seasonality) due to changes in underlying data, clinical environment, or other factors.[122,198] Though it is readily apparent when it occurs suddenly, dataset shift can occur gradually over time, where users of the model may not recognize changes in model performance. One form of dataset shift that often occurs gradually is calibration drift, in which the model gradually beings to systematically overestimate or underestimate risks.[121,134,144] Calibration drift is described as a consequence of deploying models in non-stationary clinical environments where differences in event rates arise over time between the population on which a model was developed and the population to which the model is applied.[144]

Prostate cancer is one clinical area in which national guidelines recommend the use of models to support decision-making with respect to biopsy and surgery. However, these recommendations occur despite changes in screening and treatment patterns in response to evolving evidence from randomized trials.[4,21,199] Changes in prostate cancer screening guidelines have resulted in fewer diagnosis of low-risk disease and proportionally more intermediate and high risk nationally.[200] Similarly, active surveillance rates have been increasing nationally and in Michigan,[201,202] which has led to fewer low-risk men undergoing radical prostatectomy. Given these secular trends, prediction models derived from historical registry data may be susceptible to calibration drift. Though studies based on simulation and electronic health record (EHR) data have shown how models can deteriorate over time[134,144,198], what remains unknown is how secular trends affect models trained on registry data. While EHR data have several limitations[117] in part because of their collection to support multiple needs (e.g., billing, clinical documentation,

and care delivery), registries are designed for disease surveillance and quality improvement with standardized data abstraction protocols in place with high-level of consistency in definitions and measures.[52]

Secular trends may affect model performance by systematically over- or under-predicting risks. In addition to undermining clinicians' confidence in such models, calibration drift may result in poor decision-making because clinical guidelines anchor decisions to absolute risk.[144] The purpose of this chapter is to examine the impact of secular trends on prediction model performance and to compare model updating strategies to correct for these changes. Prior research has proposed several strategies for updating or recalibrating models in the face of miscalibration. Steyerberg proposed several model updating strategies including no updates, recalibration (i.e., updating intercept and recalibration of intercept and slope), model revision (i.e., recalibrating, and selective re-estimating of regression coefficients), to model extension (i.e., re-estimating coefficients and adding new predictors).[203] Additionally, Strobl, Hickey and others proposed recalibration or retraining models at pre-specified time periods.[116,134,157,204] Similarly, Vickers et al. implemented a differential weighting scheme where greater weights are given to recent patients, who then exerted greater influence on model coefficients.[111] This approach introduces the concept of "forgetting" where data from earlier years are considered less important when updating models.[203] These studies demonstrate models need to be updated periodically as a response to dataset shift including secular trends in clinical practice patterns or other changes. In this chapter, we assess a statewide registry for secular trends in practice patterns and compare two model updating strategies in predicting continence after radical prostatectomy by comparing key model performance metrics.

5.3 Methods

5.3.1 Data Sources

We used the Michigan Urological Surgery Improvement Collaborative (MUSIC) registry to assess secular trends in practice patterns, build prediction models using the MUSIC Patient Reported Outcomes (PRO) data and assess model performance over multiple years. MUSIC registry and MUSIC-PRO have been described in previous chapters. Briefly, MUSIC collects patient reported outcomes data using the Expanded Prostate Cancer Index Composite (EPIC-26)

instrument from men undergoing radical prostatectomy, benchmarks patient reported outcomes and provides performance feedback to corresponding surgeons. MUSIC PRO is integral to improving the quality of prostate cancer care in Michigan, one of MUSIC's key priorities.

5.3.2 Study Cohort

We included men with prostate cancer who underwent radical prostatectomies between January 2016 and December 2021 and completed the MUSIC PRO questionnaire at baseline and 3-months after RP. Men who had radical prostatectomies prior to January 2016 were excluded from the analysis as MUSIC transitioned from the Memorial Sloan Kettering Cancer Center's tool to EPIC-26 in 2015-2016.[88] Though a crosswalk between these two instruments has been developed, we chose to limit our analysis to 2017 to 2021 to reduce any potential measurement errors that resulted from the transition.

5.3.3 Outcomes

Our outcome of interest is urinary pad use at 3-months after radical prostatectomy. We selected pad use at 3-months as our outcomes because it provides an early assessment of continence recovery and is also the earliest time point when EPIC-26 instrument is administered in MUSIC. Pad use is assessed in EPIC-26 as the number of pads or adult diapers a respondent might have used per day to control leakage during the last 4 weeks. Responses to the pad use question were dichotomized into continent (0 pad per day) versus incontinent (1 or more pads per day).

We assessed secular trends in prostate cancer incidence and changes in the uptake of radical prostatectomy based on the National Cancer Consortium Network (NCCN) risk groups. The NCCN risk stratifies diagnoses into five groups very low-risk, low-risk, intermediate risk (including favorable and unfavorable risks), high-risk and very high-risk.[205] For simplicity, we consolidated these 5 groups into 3 broader categories; low-risk (by combining very low and low risk groups), intermediate risk, and high-risk (by combining high-risk and very high-risk groups).

5.3.4 Predictors

Model predictors were patient's age at radical prostatectomy, body mass index, Charlson comorbidity index, Gleason grade group, clinical T-stage, pre-operative Prostate Specific Antigen (PSA), and percent positive cores at biopsy.

5.3.5 Comparison of Model Updating Strategies

We trained logistic regression models with MUSIC PRO data to predict men likely to remain incontinent at 3-months after radical prostatectomy and evaluated the performance of the models across subsequent years. As a baseline approach, we evaluated models trained on 2017 data in subsequent years with no updates to the initial model. We compared the baseline approach against two model updating strategies. In the first strategy, models were re-trained every year only using the prior year's data. In the second strategy, models were re-trained every year using all available historical data up to that year (for example, the 2019 model was re-trained on an aggregate of 2017 and 2018 data). Our model updating strategies build on the strategy implemented by Strobl et al., who employed cumulative training sets that grew each year and the validation set changed each year.[116] In the present study, we used rolling and cumulative cohort for our 2 model retraining strategies.

5.3.6 Model Performance

We assessed model discrimination and calibration by statistically and graphically evaluating how models performed each year for the baseline approach and the two model updating strategies. Discrimination refers to a model's ability to differentiate between individuals who have an event (or outcome of interest) and those who do not have an event. In the present study, discrimination refers to the ability of prediction models to differentiate between those who use pads versus those who do not. Calibration is the accuracy of risk estimates which relates to the agreement between the probability of developing the event (or, outcome of interest) as estimated by the model and the observed number of events.[126,150,206] In the present study, calibration refers to how well a model predicts the probability of being pad free with respect to

the actual probability of being pad free by assessing the closeness between a model's estimated probability and the observed probability. Calibration is assessed using both the intercept and the slope and a well-calibrated model will have an intercept of 0 and a slope of 1. Calibration intercept compares the mean observed values with mean predicted values and represents overall miscalibration, while calibration slope measures the direction of miscalibration by assessing the spread of the predicted risks.[150,153,203] The definition of calibration has evolved over time with some defining calibration as the degree to which numerical predictions are too high or too low compared to the outcomes.[155] In an ideal environment, a perfectly calibrated model will predict for all patients the same predicted risk, which equals to the event rate. Similarly, in an ideal situation, a model with perfect discrimination will always distinguish a person with an event from a person without an event.[207]

Discrimination was assessed with the area under the receiver operating characteristic curve (AUC) and model discrimination values range from 0.5 (random chance) to 1.0 (perfect discrimination). DeLong's non-parametric approach was used to assess discriminative performance by testing the differences in the AUCs of each model.[208] DeLong's test determines whether one model has a statistically significant AUC from a comparator model, hence, we compared each model to its corresponding model across strategies. Model calibration was assessed graphically using calibration plots, and statistically using calibration slopes and intercepts. A slope of 1 is considered perfect calibration, while >1 denotes underestimation of high risk and overestimation of low risk, while <1 denotes underestimation of low risk and overestimation of high risk.[153] We evaluated the magnitude of calibration drift and assessed how model performance deteriorated as practice patterns changed.

5.3.7 Software

Secular trends in MUSIC registry data are assessed using Joinpoint linear regression. Joinpoint regression program detects statistically significant changes in secular trends in cancer surveillance research and assesses secular trends by fitting the simplest model to describe the trend data.[152] The changes are expressed as annual percentage change (APC) which is used to compare year over year trends. Logistic models were fit in R 4.1.2 (R Foundation, Vienna,

Austria). Differences in AUC curves were tested using *pROC* package while calibration intercept and slope were assessed using the *val.prob.ci.2* packages.[150,209]

5.4 Results

5.4.1 Cohort Characteristics

This study included data on 6,596 men in the MUSIC registry who underwent radical prostatectomy between January 2017 and December 2021 and completed at least one EPIC-26 survey reporting their postoperative pad use. Patient characteristics at baseline for each year are presented in Table 1. Briefly, the median age at radical prostatectomy was 64 (IQR 59,68) in 2017, which remained steady in the intervening years and increased to 65 (IQR 60, 69) in 2021. Rates of Gleason grade group 2 disease increased from 45% in 2017 to 49% and 48% in 2020 and 2021, respectively. Similarly, rates of grade group 3 disease also increased from 20% in 2017 to 26% in 2021. The rates of men with NCCN low risk disease undergoing radical prostatectomy decreased from 16% in 2017 to 7% in 2021. Similarly, men with intermediate risk disease undergoing radical prostatectomy increased from 75% in 2017 (n=1133) to 84% in 2021 (n=417). However, the rates of high risk men undergoing RP remained stable around 9% to 10% throughout the five year study period.

5.4.2 Prevalence of Outcomes

In 2017, 71% (n=898) of men in our data experienced incontinence. The prevalence of incontinence increased to 74% (n=1009) and 73% (n=1034) in 2018 and 2019 respectively, and subsequently decreased to 69% in 2020 (n=571) and 2021 (n=231).

5.4.3 Statewide Practice Patterns in MUSIC

Analyses of practice patterns in MUSIC showed active surveillance increased from approximately 32% in 2017 to 39% in 2021 (Figure 1) with an annual percent change of 5%. The rates of radical prostatectomy steadily declined in MUSIC from 43% in 2016 to about 32% in

2021 with an annual percent change of -5.66 (Figure 2). The incidence of NCCN low-risk disease declined from 30% in 2017 to 26% in 2021 with an APC of -3.92 (Figure 3). The rates of NCCN intermediate disease slightly increased over time from 47% in 2015 to 53% in 2021 with an APC of 1.92, while the rates of NCCN high-risk disease remained stable at around 17% with an APC of 0.91 (Figures 4 and 5).

5.5 Model Performance in the Validation Cohorts

Model discrimination was modest across all models predicting 3-month pad use. Overall, the AUC was stable for different model updating strategies with values ranging from 0.59 to 0.63.

5.5.1 Baseline approach

Model discrimination remained modest for the model trained at all time points in predicting pad use at month 3. The 2018 model predicted pad use with an area under the curve of 0.59 (0.56 - 0.63). The AUC remained stable at 0.62 (0.55 – 0.68) when the model was tested separately in subsequent years. Calibration was poor for the 2018 model with an intercept of 0.11 (-0.02 – 0.23) and a slope of 0.65 (0.40 – 0.90). However, model calibration improved in 2019 and 2020, with decreasing intercept and increasing slope values up until the 2021 model, which had an intercept of -0.19 (-0.43 – 0.05) and slope of 0.92 (0.40 – 1.44) indicating over-estimation of risks as a perfectly calibrated model will have an intercept of 0 and a slope of 1.

5.5.2 Model Updating Strategy 1

All four models trained with previous year's data yielded similar AUCs. The 2019 model trained on the previous year data had an AUC of 0.61 (0.58 – 0.64). The AUC remained stable to 0.62 (0.57 – 0.66) for the 2020 model and remained unchanged for the 2021 model. Calibration intercept decreased while slope increased over time. The 2020 model showed an intercept of -0.19 (-0.34 – -0.04) and a slope of 1.06 (0.67 – 1.46) and continued to overestimate risks with an intercept of -0.04 (-0.28 – 0.20) and slope of 1.08 (0.51 – 1.64) for the 2021 model.

5.5.3 Model Updating Strategy 2

All four models trained with cumulative data yielded similar AUCs as in the previous models. Calibration improved from an intercept of 0.11 (-0.02 – 0.23) and slope of 0.65 (0.40 – 0.90) for the 2018 model to an intercept of 0.01 (-0.11 – 0.13) and the slope of 0.88 (0.60 – 1.17) for the 2019 model. The subsequent models over-estimated risks with an intercept of -0.19 for 2020 and 2021 models. The slopes also increased to 1.05 and 1.25 in 2020 and 2021 respectively indicating the models yielded more extreme predictions. (Table 2)

5.5.4 Comparison of Model Updating Strategies

We compared the area under the receiver operating curve using the DeLong test to compare the two model updating strategies against the baseline approach. The DeLong tests showed no statistically significant differences between models' discriminative performance (Table 3). Assessment of calibration intercept and slope showed calibration remained stable across the three model updating approaches. When comparing the two strategies against the baseline approach, the 2019 model in strategy 1 and the 2020 model in strategy 2 were mis-calibrated. Calibration intercepts for the rest of the models showed no signs of miscalibration except the intercept values decreased over time suggesting some over-estimation. Similarly, the slope values increased from 2018 model to 2021 for baseline as well as for strategies 1 and 2 models. However, in the baseline approach the calibration slope was less than 1 for all four years (Table 3), whereas, in strategies 1 and 2, the calibration slope was greater than 1 for most years.

5.6 Discussion

In this study, we evaluated the performance of prediction models to assess if model discrimination and calibration deteriorated over time. Discrimination remained modest and stable across all models. Neither strategy produced a model with statistically significant discriminative performance when tested against the baseline approach. Overall, model calibration assessed with intercept and slope remained similar across the strategies with some indication of over-

estimation as reflected by decreasing intercept values in later years. Our model updating strategies build on the work advanced by Strobl et al.[116] Their models were trained on a cumulative sample with data from previous years included in the training set. The models were then tested on validation sets that changed each year. We adopted a similar strategy (herein, strategy 2) and added another strategy (herein, strategy 1) where we trained the models on prior year data and tested on subsequent year's data. We also compared the performance of these two strategies against the baseline approach.

In the baseline approach, the intercept decreased while slope increased over time, reflecting over-estimation of risks in later years. However, the slope values remained less than 1 indicating predictions were likely too extreme, i.e., underestimation of low risk and overestimation of high risk.[153,207] Similarly, strategy 1 models also produced over-estimated risk probabilities in later years, but the slope values were greater than 1 indicating predictions were too moderate (or narrow).[126,210] Strategy 2 models did not achieve superior calibration compared with the baseline approach.

Although the slope values for later year models were >1 in strategy 1 and 2, it is interesting to compare the intercepts and slopes against the event rates. When the event rates were higher in the training set than the testing set, the models over-estimated risks (as evidenced by the decreasing intercept values) but the risk estimates did not seem to be too extreme (as evidenced by calibration slope >1). However, a closer inspection of the confidence intervals for the slope suggests that there are no significant differences in slopes (i.e., confidence intervals cross 1, as the optimal value of slope is 1) when the event rates for the testing year were smaller than for the training year.

The present study showed how model discrimination and calibration remained stable over time, albeit with increasing over-estimation of risk in later years. Our results of stable discriminative performance and increasing overestimation of risks are similar to other studies, [116,134,198] except we did not detect strong evidence of calibration drift. Visual inspection of calibration plots revealed no clear evidence of calibration drift in segments of the curves with most of the observations. Areas of over- and under-estimations were primarily observed in areas with few observations (i.e., predicted probabilities). Hence, this study generates a hypothesis that models trained on registry data may be less susceptible to calibration drift. As discussed above,

registry data tend to be stable with standardized data abstraction protocols, which may explain why we did not observe calibration drift despite changing practice patterns.

This study has several limitations. First, our sample size in annual cohorts was small and fewer men underwent radical prostatectomy during the first two years of COVID, further reducing the cohort size in later years. This may have reduced power to assess calibration drift over time, however, assessment of statistical power in calibration studies is an under-studied area.[150] Second, we used 5 years of data to build and retrain our models and the temporal change during these years may not have been profound to cause calibration drift. It is reasonable to assume that a longer time-frame may have shown evidence of calibration drift. Third, model discrimination was modest across all models, which may have affected model calibration. It is likely that models with poor discrimination may also have poor calibration, however, previous analyses have shown the difficulties of predicting 3 month outcome from pre-operative variables. Fourth, we could not ascertain if the decrease in event rates were permanent or transitory. The decreasing event rates could have been an artifact of temporary changes in practice patterns due to COVID in 2020 and 2021, which could self-correct over time as COVID evolves into an endemic stage.

Despite the limitations, the present study has several implications for prediction models trained on registry data. First, registry data may be less prone to shifts that are emblematic of EHR data. Therefore, calibration drift solutions devised for EHR data may not necessarily apply to models trained on registry data. This study focuses on simple model re-training efforts and compares model performance against the baseline approach to assess discrimination and calibration. Other disease registries may find our experience instructive. Despite the lack of clear evidence of calibration drift, this study adds to the limited literature on model calibration in registry data. With the expanding use of registry data and the move towards sophisticated prediction models, it is reasonable to expect that future prostate cancer patients would benefit from greater use of risk predictions models trained on registry data. This advancement is likely to positively impact future patients. However, a poorly calibrated model will consistently assign misleading risk probabilities. Therefore, as more prediction models are being used in routine clinical practice, model surveillance, maintenance, and updating become critical to safe and effective patient care.[122,203,211] Future studies should confirm our findings with a larger sample size and assess model performance over a long time horizon.

5.7 Conclusion

In this study, we assessed if models deteriorate over time as practice patterns changed in a registry data. Robustness of predictions is important to the broader use of prediction models in clinical settings. As clinicians treating localized prostate cancer patients strive to balance the tradeoffs between potential cure with treatment side effects, assigning accurate probability of optimal outcomes on a given therapy is critical in the treatment decision-making process.

Table 10. Characteristics of Study Participants Stratified by Testing Years

Characteristic	2017 N = 1,513	2018 N = 1,605	2019 N = 1,746	2020 N = 1,232	2021 N = 500
Age (Median, IQR)	64 (59, 68)	64 (59, 68)	64 (59, 68)	64 (60, 69)	65 (60, 69)
BMI (kg/m2, n(%))					
<25	250 (17%)	260 (16%)	258 (15%)	195 (16%)	58 (12%)
25-30	657 (44%)	678 (43%)	764 (44%)	545 (45%)	242 (49%)
>30	590 (39%)	655 (41%)	711 (41%)	465 (39%)	194 (39%)
Race					
Black	156 (10%)	170 (11%)	178 (10%)	117 (9.5%)	58 (12%)
White	1,115 (74%)	1,209 (75%)	1,308 (75%)	965 (78%)	371 (74%)
Other	32 (2.1%)	30 (1.9%)	36 (2.1%)	24 (1.9%)	18 (3.6%)
Unknown	208 (14%)	195 (12%)	223 (13%)	126 (10%)	53 (11%)
Charlson Comorbidity Index					
0-1	1,354 (90%)	1,410 (88%)	1,562 (89%)	1,125 (91%)	456 (91%)
2+	156 (10%)	194 (12%)	184 (11%)	107 (8.7%)	44 (8.8%)
PSA					
<4	1,098 (73%)	1,178 (73%)	1,271 (73%)	935 (76%)	389 (78%)
4-10	310 (20%)	294 (18%)	300 (17%)	213 (17%)	84 (17%)
>10	105 (6.9%)	132 (8.2%)	174 (10.0%)	81 (6.6%)	27 (5.4%)
Clinical T-Stage					
T1	1,107 (73%)	1,195 (75%)	1,336 (77%)	973 (79%)	391 (78%)
T2a	194 (13%)	171 (11%)	173 (9.9%)	113 (9.2%)	52 (10%)
>T2a	208 (14%)	235 (15%)	234 (13%)	144 (12%)	56 (11%)
Gleason Grade Group					
1 (3+3)	263 (18%)	212 (13%)	254 (15%)	144 (12%)	46 (9.3%)
2 (3+4)	671 (45%)	716 (45%)	787 (46%)	596 (49%)	239 (48%)
3 (4+3)	303 (20%)	384 (24%)	394 (23%)	314 (26%)	131 (26%)
4 (4+4)	154 (10%)	151 (9.5%)	186 (11%)	108 (8.8%)	56 (11%)
5 (>4 + >4)	106 (7.1%)	128 (8.0%)	106 (6.1%)	63 (5.1%)	23 (4.6%)
NCCN Risk Groups					
High Risk	131 (8.7%)	148 (9.3%)	166 (9.6%)	93 (7.6%)	46 (9.2%)
Intermediate Risk	1,133 (75%)	1,262 (79%)	1,351 (78%)	1,005 (82%)	417 (84%)
Low Risk	240 (16%)	180 (11%)	220 (13%)	127 (10%)	35 (7.0%)

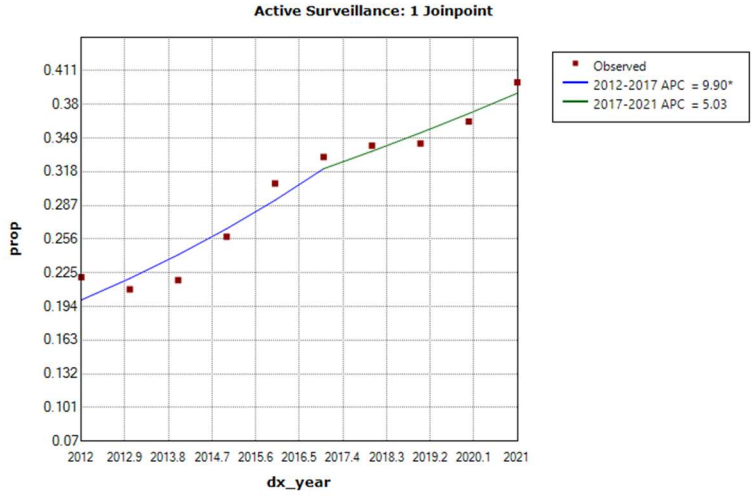
Table 11. Model Performance by Modeling Updating Strategy

Year		Event Rate		Intercept	Slope	AUC
Training	*Testing*	*Training*	*Testing*	*(CI)*	*(CI)*	*(CI)*
Baseline Approach						
2017	2018	0.71	0.74	0.11 *(-0.02 – 0.23)*	0.65 *(0.40 – 0.90)*	0.59 *(0.56 – 0.63)*
2017	2019	0.71	0.73	0.06 *(-0.07 – 0.18)*	0.72 *(0.49 – 0.96)*	0.61 *(0.58 – 0.64)*
2017	2020	0.71	0.69	-0.14 (*-0.29 – 0.02*)	0.77 *(0.45 – 1.09)*	0.60 *(0.56 – 0.64)*
2017	2021	0.71	0.69	-0.19 *(-0.43 – 0.05)*	0.92 *(0.40 – 1.44)*	0.62 *(0.55 – 0.68)*
Strategy 1						
2017	2018	0.71	0.74	0.11 *(-0.02 – 0.23)*	0.65 *(0.40 – 0.90)*	0.59 *(0.56 – 0.63)*
2018	2019	0.74	0.73	-0.05 *(-0.17 – 0.07)*	1.11 *(0.76 – 1.46)*	0.61 *(0.58 – 0.64)*
2019	2020	0.73	0.69	-0.19 *(-0.34 – -0.04)*	1.06 *(0.67 – 1.46)*	0.62 *(0.57 – 0.66)*
2020	2021	0.69	0.69	-0.04 *(-0.28 – 0.20)*	1.08 *(0.51 – 1.64)*	0.62 *(0.56 – 0.69)*
Strategy 2						
2017	2018	0.71	0.74	0.11 *(-0.02 – 0.23)*	0.65 *(0.40 – 0.90)*	0.59 *(0.56 – 0.63)*
2017 - 2018	2019	0.73	0.73	0.01 *(-0.11 – 0.13)*	0.88 *(0.60 – 1.17)*	0.61 *(0.58 – 0.64)*
2017 - 2019	2020	0.73	0.69	-0.19 (*-0.34 – -0.03*)	1.05 *(0.65 – 1.45)*	0.61 *(0.57 – 0.65)*
2017 - 2020	2021	0.72	0.69	-0.19 *(-0.43 – 0.05)*	1.25 *(0.59 – 1.90)*	0.62 *(0.56 – 0.69)*

Table 12. ROC Test Comparison by Model Updating Strategy

Year	AUC (Baseline)	AUC (Strategy 1)	Z-Statistic	Lower CI	Upper CI	p-value
2018	0.59	0.59	0.00	0.000	0.000	1.000
2019	0.61	0.61	-0.15	-0.010	0.008	0.884
2020	0.60	0.62	-1.29	-0.033	0.007	0.197
2021	0.62	0.62	-0.14	-0.044	0.038	0.890
	AUC (Baseline)	**AUC (Strategy 2)**				
2018	0.59	0.59	0.00	0.000	0.000	1.000
2019	0.61	0.61	-0.25	-0.004	0.004	0.805
2020	0.60	0.61	-1.62	-0.017	0.001	0.106
2021	0.62	0.63	-0.52	-0.025	0.007	0.606

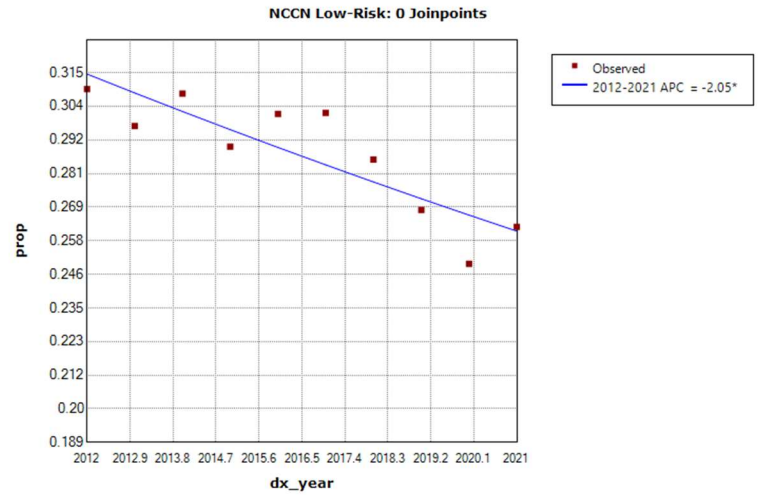
Figure 9. Changes in Active Surveillance in MUSIC Registry 2012 – 2021



* Indicates that the Annual Percent Change (APC) is significantly different from zero at the alpha = 0.05 level
Final Selected Model: 0 Joinpoints.

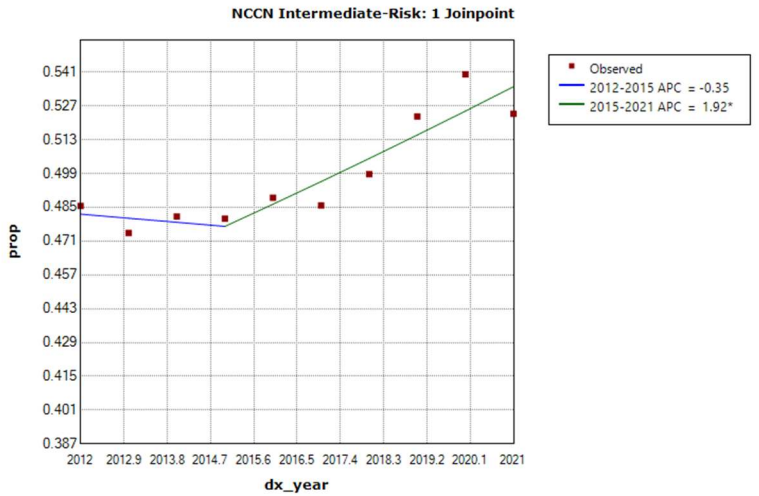
Figure 10. Changes in Radical Prostatectomy in MUSIC Registry 2012 – 2021

Figure 11. Changes in NCCN Low Risk Cancers in MUSIC Registry 2012-2021



* Indicates that the Annual Percent Change (APC) is significantly different from zero at the alpha = 0.05 level
Final Selected Model: 0 Joinpoints.

Figure 12. Changes in NCCN Intermediate Risk Prostate Cancers in MUSIC Registry 2012-2021



* Indicates that the Annual Percent Change (APC) is significantly different from zero at the alpha = 0.05 level
Final Selected Model: 0 Joinpoints.

Figure 13. Calibration Plot, Baseline Approach 2018

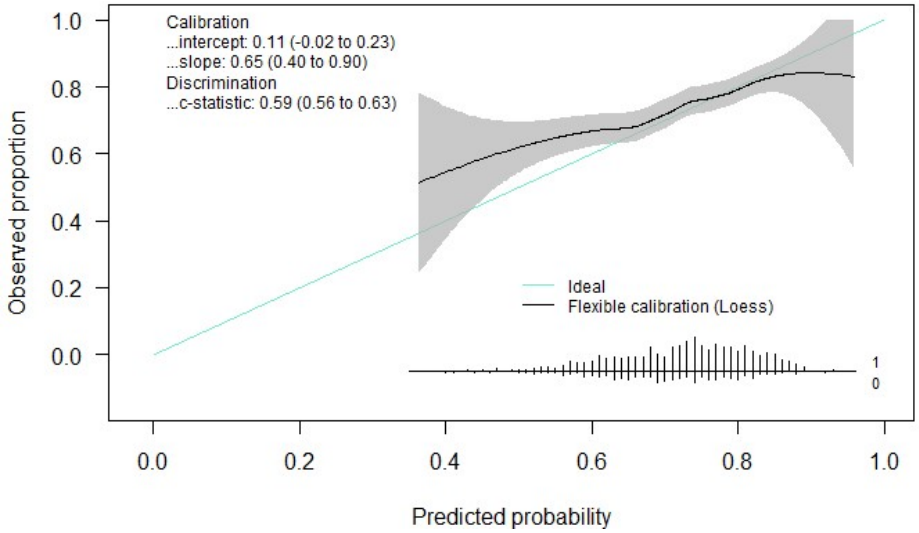


Figure 14. Calibration Plot, Baseline Approach 2019

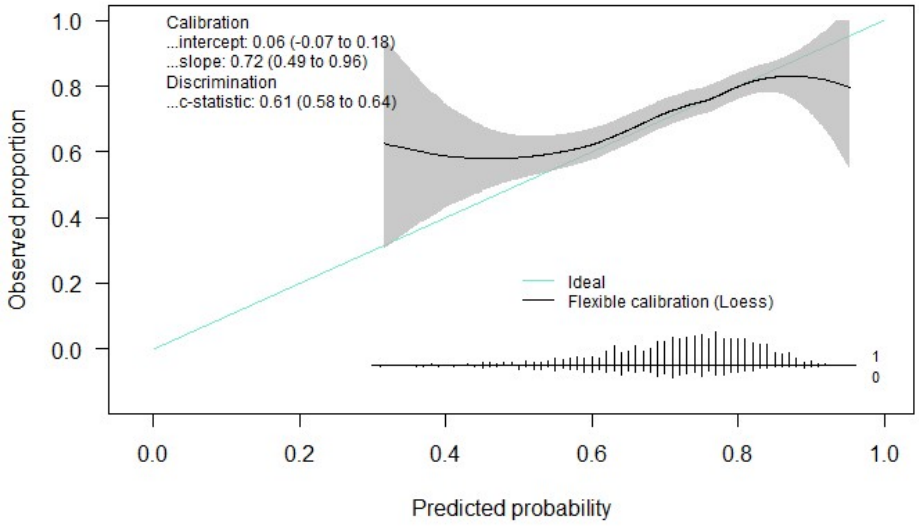


Figure 15. Calibration Plot, Baseline Approach 2020

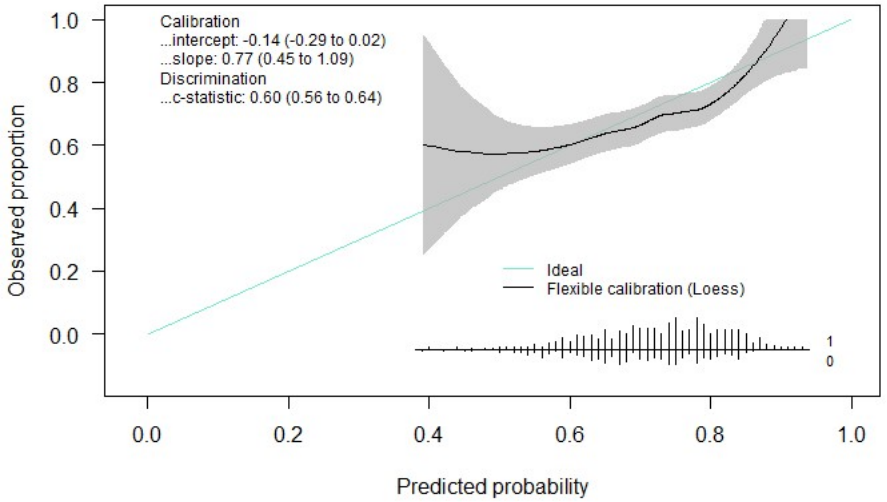


Figure 16. Calibration Plot, Baseline Approach 2021

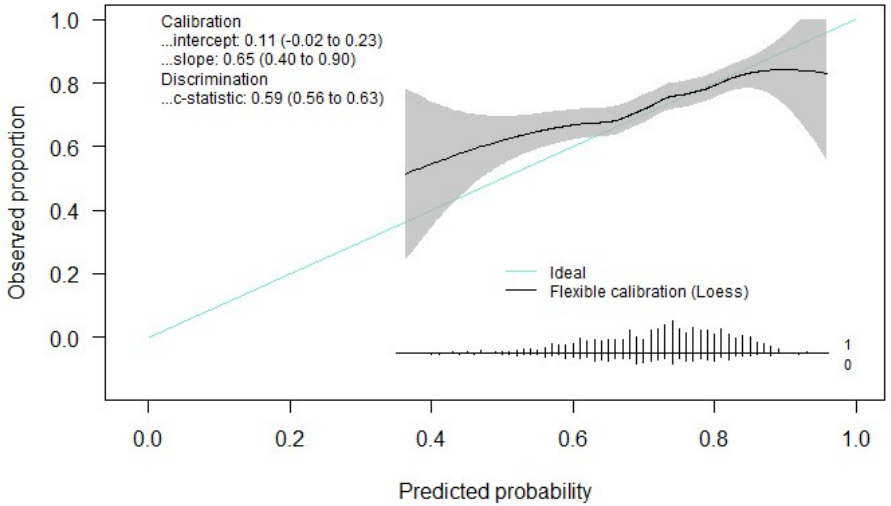


Figure 17. Calibration Plot, Strategy 1, 2018

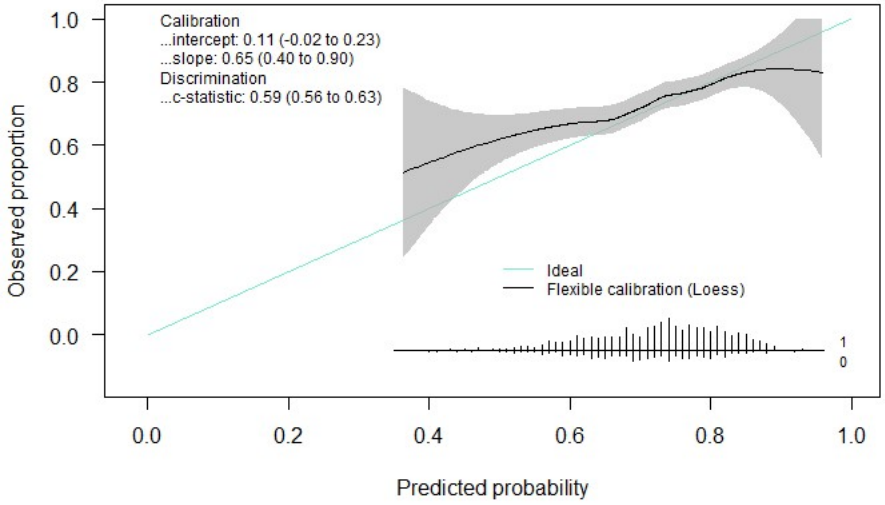


Figure 18. Calibration Plot, Strategy 1, 2019

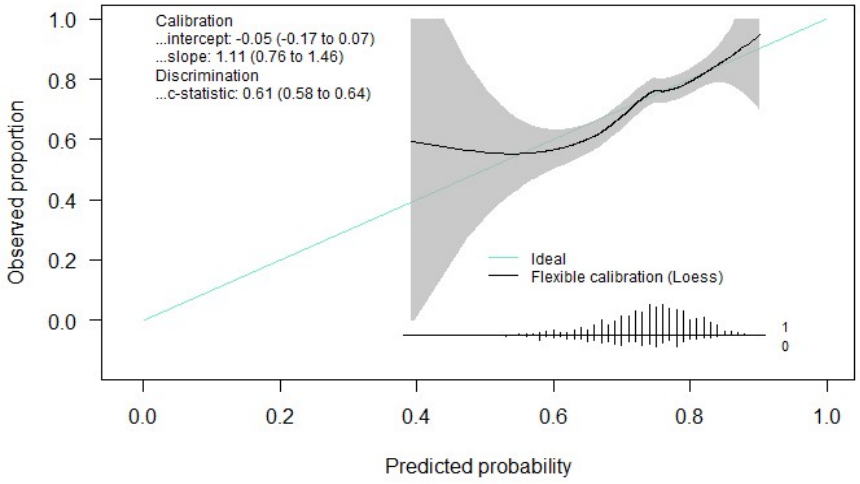


Figure 19. Calibration Plot, Strategy 1, 2020

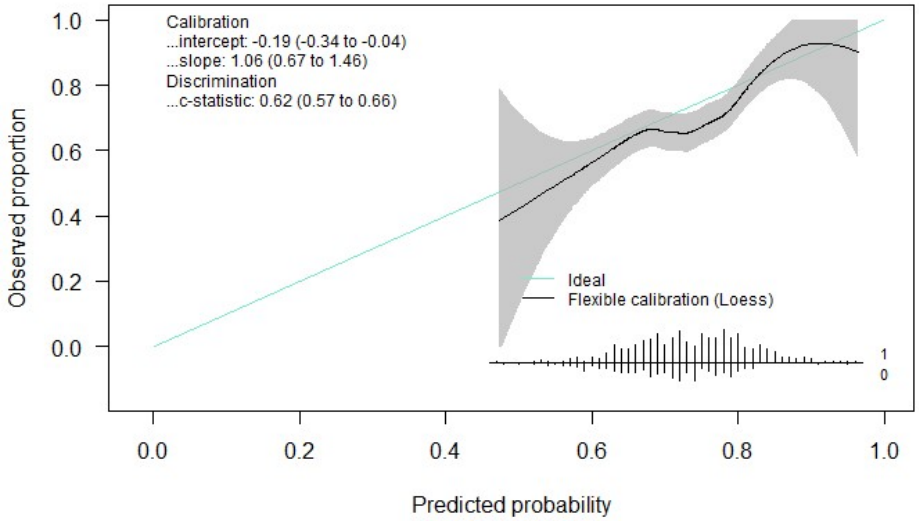


Figure 20. Calibration Plot, Strategy 1, 2021

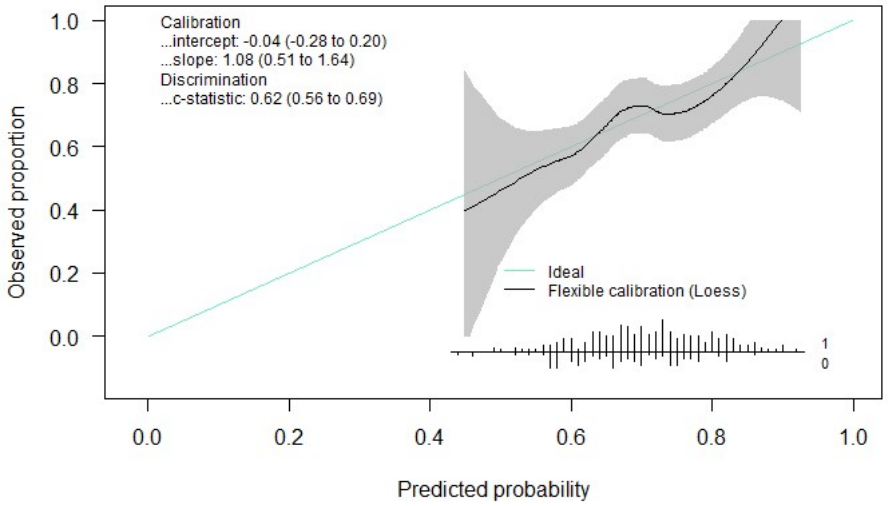


Figure 21. Calibration Plot, Strategy 2, 2018

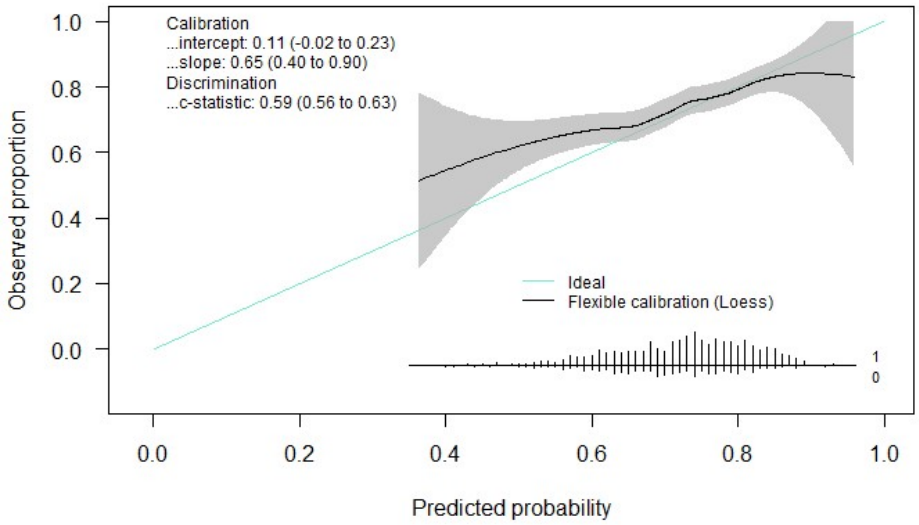


Figure 22. Calibration Plot, Strategy 2, 2019

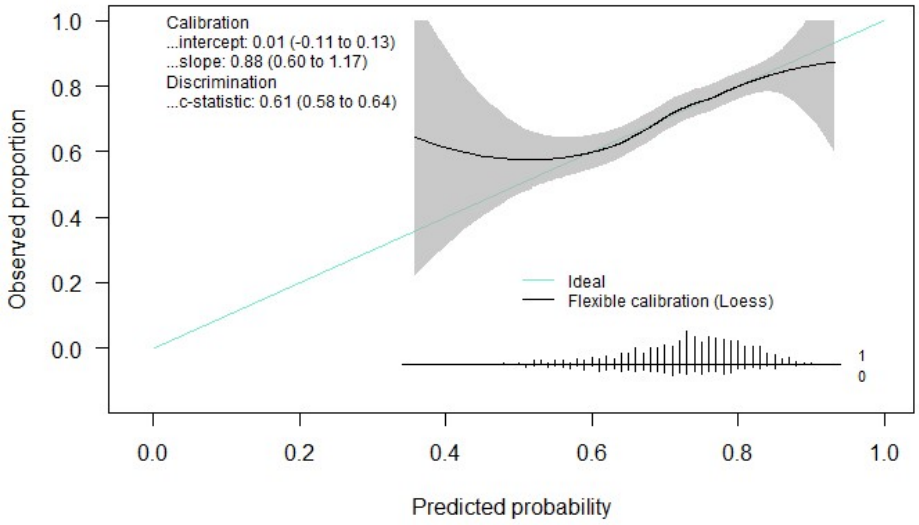


Figure 23. Calibration Plot, Strategy 2, 2020

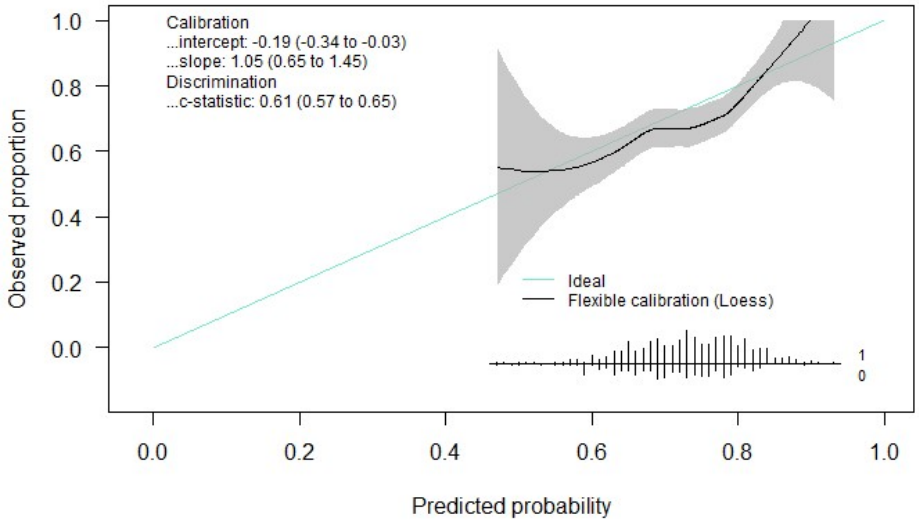
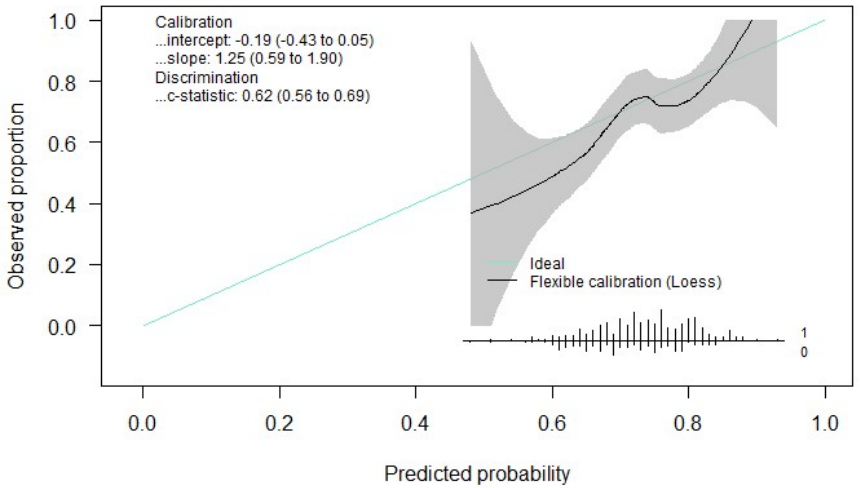


Figure 24. Calibration Plot, Strategy 2, 2021



Chapter 6 Discussion

Randomized trials show radical prostatectomies prolong life but also affect the patient's quality of life including urinary incontinence. Patient Reported Outcomes (PROs) are used to assess the post-surgical quality of life. Longitudinal analyses of quality of life based on PROs is the most patient-centered way of exploring functional recovery after radical prostatectomy. PROs also play a critical role in quality improvement efforts as they bring the patient's voice to the table. Investigation of patient reported outcomes is important because of the potential to learn from patients in routine clinical practice, the idea behind the concept of a learning health system. Systematic investigation of long-term functional recovery builds real-world evidence to support treatment decision-making process for future patients.

6.1 Implications for Patient Care

This dissertation explores longitudinal aspects of patient reported outcomes with a focus on urinary incontinence recovery. This dissertation makes several scientific contributions to the literature, focusing on the use of PROs to broaden our understanding of continence recovery after radical prostatectomy. In chapter 3, we explored variability in urinary function recovery using mixed-effects models to characterize the patterns of variability. We found urinary function recovery is highly variable among patients who undergo radical prostatectomy. The amount of variability attributable to patients was much greater than attributed to surgeons. These findings help to contextualize factors that impact urinary function recovery and adds to our understanding of the recovery process in a patient-centered manner. First, PROs are unfiltered assessment of recovery from patient's perspective. They are an important aspect of recovery and bring the patient's voice to the table. Further, assessing factors that explain the variability in urinary function recovery has implications for quality improvement efforts– i.e., to explore how much of the variability is due to modifiable factors amenable to quality improvement efforts. Factors that are modifiable can be addressed through quality improvement efforts and can help improve

patient outcomes. Identifying patients who are at risk for late recovery might shorten the duration of post-surgical incontinence and improving their quality of life. Hence, urology practices can assess continence at an earlier time point, so patients do not have to wait for 3 months, which is customary for many practices. Offering early interventions to help regain urinary function sooner has enormous quality of life benefits for patients.

We also assessed if surgical volume (or surgeon's case-load) is associated with urinary function recovery. Surgeon's case-load did not fully explain the variability in outcomes. This finding makes an important contribution to our understanding of volume-functional outcomes relationship as surgeon's case-load is an important aspect of patient care. High volume surgeons and high volume centers have better clinical and oncologic outcomes. However, studies have shown no measurable improvements in urinary function outcomes over time even among high volume surgeons. Also, MUSIC experience suggests that continence outcomes have not improved despite the concerted quality improvement efforts. This raises an important question about where to direct quality improvement efforts. Should quality improvement efforts be targeted at identifying patients who are at risk for later recovery? Should patients be offered continence interventions earlier in the recovery process to improve their quality of life? The subsequent chapters attempt to provide some answers to these questions.

In chapter 4, we explored the concept of dynamic modeling to predict urinary continence at multiple time points after radical prostatectomy. In this analysis, we trained models to longitudinally predict pad use and urinary function by incorporating patient reported data that become available postoperatively. This approach was based on the hypothesis that updating predictions as new data become available would make postoperative models more accurate and responsive to the recovery patterns. We found models trained on both baseline and post-operative data produced better estimates of urinary continence and function. The finding that post-operative models predicted urinary function recovery better than the preoperative models may not be surprising but the idea of adding post-operative data to prediction models is a relatively new concept in prostate cancer care. This is an important contribution to the literature.

Traditional nomograms are static and are not responsive to longitudinal recovery trajectory. By demonstrating that postoperative models perform better at predicting urinary continence and functional recovery than preoperative models, this work supports the use of postoperative models in clinical practice. Since functional recovery is longitudinal and dynamic,

it would serve clinicians and patients well if prediction models are more responsive to the recovery trajectory. If patients recover slower than initially expected, post-operative predictions can be used to reassure patients, reset their expectations or to recommend rehabilitative or other interventions. If models can identify patients who are less likely to achieve continence or recover urinary function at an earlier time point, it would help clinicians recommend rehabilitative or other interventions. This would improve the quality of life by offering early interventions and potentially helping the patients become continent earlier in their survivorship journey.

We also found that postoperative models are better at estimating short-term continence outcomes than predicting outcomes over a longer time horizon (i.e., the 3 month model predicting 6 month outcomes and the 6 month model predicting 12 month outcomes). Based on this finding, we can hypothesize that a 1 month prediction model will provide similar estimates as the 3 month model. The potential to estimates recovery outcomes at 1 month after surgery has great implications for patient care. MUSIC has started collecting PROs within the first month after surgery, which could help identify patients who are at higher risk of late continence recovery before the traditional 3 month time-frame. Future modeling efforts should address if a prediction model trained on 1 month data performs better than the models trained on months 3 or 6 data.

In chapter 5, we evaluated the performance of prediction models longitudinally. We trained logistic regression models to predict the probability of incontinence at 3 months after radical prostatectomy. We assessed temporal changes in practice patterns in MUSIC and assessed how those changes may affect the performance of prediction models over time. We compared the baseline approach with no model retraining, against two model updating strategies. The baseline approach is the most prevalent approach as prediction models in urology rarely get updated over time. We found that model discrimination remained modest and stable across the models. Other studies have shown temporal changes do not impact model discrimination as much as they impact model calibration.

We assessed the extent of miscalibration by examining changes in model intercepts and slopes. We found model calibration remained stable and there was no clear indication of model deterioration. Prediction model developers have known that models trained on EHR data experience calibration drift over time. By showing that models trained on registry data did not experience calibration drift as practice patterns changed, this study advances our understanding

in a couple of ways. First, it generates a hypothesis that models trained on registry data may be more stable than models trained on EHR data. Second, our findings from Chapter 4 suggest that prediction models trained on registry data perform better if postoperative data are included in the models. Findings from Chapter 5 suggest that once the models are trained, they may not need to undergo periodic retraining as the underlying data tend to be more stable in clinical registries than in the EHR.

6.2 Future Directions

There are several opportunities for further research. First, there is no agreement on the definition of continence in the literature. Some have argued for a conservative definition of 0 pads per day as the most patient-centered definition. Others have adopted social continence, 1 or less pads per day, as their definition of continence. Different definitions have implications for assessing variability as well as training prediction models. A stricter definition of continence helps to assess the true variability in continence recovery and make accurate predictions. Future studies could build more robust evidence by assessing variability in continence recovery using the stricter 0 pads per days and building prediction models based on a consistent definition of continence.

Second, we were not able to show a strong association between surgeon's caseload and urinary function recovery. The association between surgeon's caseload and urinary function recovery is an important question for continued research. Surgical skills are critical to performing complete nerve spare and to preserve the underlying pelvic musculature, which helps with sexual function recovery. However, this may not extend to urinary continence recovery. Further, anatomical data such as membranous urethral length are not routinely collected in MUSIC registry. Therefore, we were not able to account for patients' anatomical differences in our studies. Future studies should investigate variability in continence recovery by incorporating pertinent anatomical data collected at biopsy and surgery. Also, future research should assess whether anatomical data further improve the performance of preoperative and postoperative prediction models.

Third, we do not know the minimum number of PRO questionnaires required for each surgeon to reliably estimate the differences in patient reported outcomes between high and low

volume surgeons. Our results show that about 65% of the surgeries in our cohort are performed by high volume surgeons and only 13% of surgeries were performed by low and mid-low volume surgeons. Our estimates are likely to be more precise for high volume surgeons and less precise for low volume surgeons. Hence, future studies should investigate the minimum number of PRO-questionnaires required to reliably assess the quality of care based on patient reported outcomes.

Lastly, there is an opportunity to investigate the performance of models predicting early continence recovery. Traditionally, EPIC-26 is administered at 3 months after surgery. Recently, MUSIC began collecting PROs at 1 month after surgery, which provides an earlier assessment of continence recovery. Prediction models based on month 1 outcomes may identify men who could benefit from early interventions. If prediction models can help shorten the time between surgery and early continence intervention, patient's quality of life is likely to improve substantially. Future work should focus on learning from all patients to improve the quality of life. The Learning Health System has the potential to provide a platform to learn from all patients by exploring what works in prostate cancer care, understanding why it works, and for whom does it work to improve overall outcomes for all patients.

6.1 Implications: Learning from Patients in a Learning Health System

Prostate cancer is a common disease, yet it is also a clinically complex, heterogeneous disease. Each prostate cancer patient presents with a complex set of biological, environmental, and epigenetic differences and each patient responds differently to treatment(s). Over and under-treatment of prostate cancer are widely recognized problems as variability in tumor biology and patient factors make it challenging for clinicians to anticipate how a patient might do and how soon a patient might regain function after clinical interventions. Therefore, prostate cancer care is an appropriate focus for a Learning Health System (LHS) approach. An LHS approach provides a framework for continuous learning through integration of research activities and clinical practice and provides an opportunity to learn directly from patients through patient reported outcomes.

The National Academies of Medicine envisioned an LHS ". . . in which progress in science, informatics, and care culture align to generate new knowledge as an ongoing, natural

by-product of the care experience, and seamlessly refine and deliver best practices for continuous improvement in health and healthcare."[212] An LHS approach provides a platform to transform how evidence is generated, aggregated, synthesized, disseminated and implemented. In an LHS, internal data and experience are systematically integrated with external evidence and that knowledge is put into practice.[213] The aspirational vision of LHS proposed by Friedman et. al. advances cyclical learning that converts real-world data captured from clinical encounters into new knowledge to continuously inform and improve clinical practice.[128] The concept of LHS assumes that the capture of practice changes and consequences of these changes generate new data, complete the cycle and initiate subsequent iterations of the learning cycle to continue to identify best practices and improve outcomes.[214]

A growing interest to include PROs in quality improvement and value-based reimbursement have driven some policy changes at the national level. The Centers for Medicare and Medicaid Services (CMS) has proposed a gradual implementation of PROs for symptom management in oncology, reflecting the understanding that PROs should be included in policy and reimbursement discussions. Policy must care about patient experience and outcomes and the best mechanism to systematically collect the information is directly from patients in the form of PROs. However, systematically collecting PROs and learning from patients has yet to be part of a unifying agenda. Some high volume institutions have advanced patient-centered research by focusing on PROS but only a few concerted efforts to collect PROs currently exist. MUSIC is an exception and a leader in this space.

6.2 MUSIC as a Learning Health System Infrastructure

MUSIC is a physician-led quality improvement organization that strives to improve quality of prostate cancer care activities by collecting clinically relevant data, comparing performance among urologists, and sharing best practices across the state of Michigan. MUSIC was founded in 2011 with support from Blue Cross and Blue Shield of Michigan.[52] As a continuous quality improvement organization, MUSIC is an infrastructure built on a set of strong governing principles to support statewide improvement in outcomes related to prostate and kidney diseases. MUSIC has developed data governance structures to collect data on urology patients statewide and maintains a secure, web-based statewide urological care registry in

Michigan.[215] MUSIC collects patient demographics, cancer characteristics including pathological details from biopsies and confirmatory testing, health care utilization and outcomes for urologic care.[216,217] MUSIC collaborative-wide data are available for analysis and quality improvement activities[9] and MUSIC offers a variety of mechanisms for continuous learning and skill development for its member clinicians including video training, publications of findings from real-world data, administration of clinical trials to test novel hypothesis, devices or techniques among other learning activities.

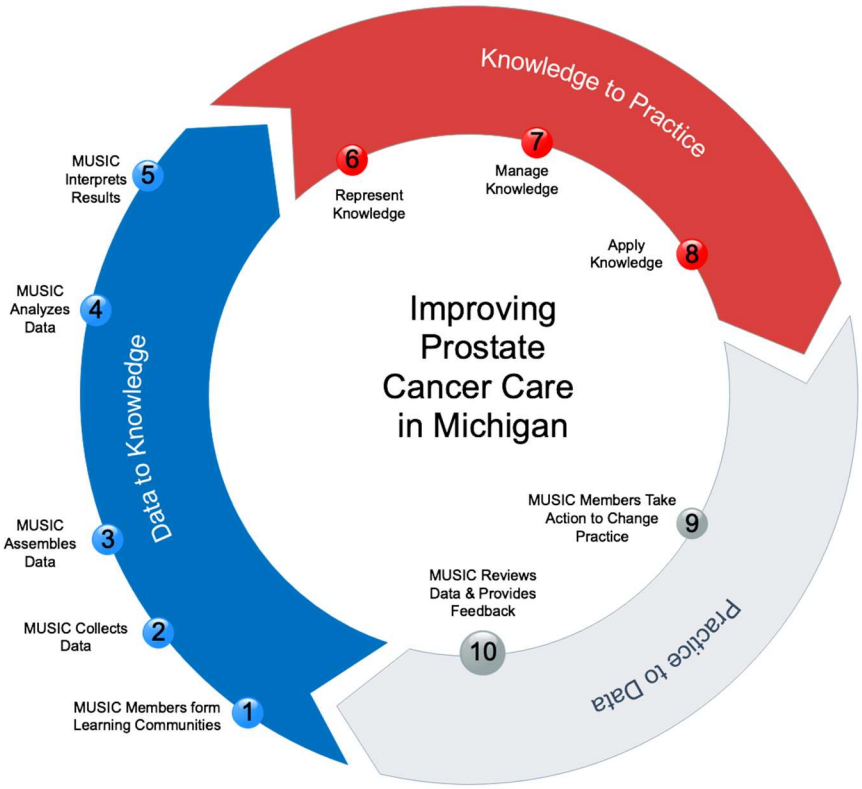
MUSIC leaders have argued that variations in prostate cancer care is inevitable because most prostate cancer patients with localized disease fall in the "grey areas" of clinical decision-making. They proposed that regional cooperatives, such as MUSIC, can reduce variations in outcomes by facilitating comparative feedback to clinicians on their patterns of care relative to their peers and existing guidelines.[218] As prostate cancer care becomes more personalized, an LHS approach help to further sort out these "grey areas." Clinicians who treat prostate cancers routinely confront difficult questions: does a patient need to be actively treated for his prostate cancer? When does the patient need to be treated, how much, and in what order a treatment gives the patient the most optimal outcome? Randomized trials could provide definitive answers to these questions but may not be able to sort out all the "grey areas." An LHS approach can help to provide a framework to answer these questions and build evidence from real-world data. In a learning health system, an iterative cycle begins with conversion of data to knowledge and when this knowledge is applied in routine clinical practice, practice changes could improve quality of care. The iterative cycle continues until desired outcomes are achieved.

MUSIC combines clinical care and quality improvement. At a very basic level, the iterative cycle is a bi-directional feedback loop where data collection is embedded into care delivery processes and care is changed in response to the evidence generated.[219] Figure 11 illustrates how MUSIC combines these aspects while using the learning cycle to improve outcomes in Michigan.

1. MUSIC receives clinical encounter and patient reported outcomes data
2. Data are systematically collated and analyzed by MUSIC
3. New knowledge is generated in the form of aggregated PRO reports, findings are published in peer-reviewed journals and clinicians are educated

4. MUSIC practices implement new knowledge by making changes to care delivery processes and improving outcomes over time
5. Patients receive evidence-based care as a result of the implementation of new knowledge
6. Cyclical learning process continues as new knowledge is generated to improve patient outcomes

Figure 25. Learning Health Cycle Conceptualization with MUSIC Activities



By performing these activities, MUSIC has established itself as a strong statewide quality improvement collaborative with its clinicians serving as a members of learning community. MUSIC is in a strong place to continue providing quality improvement support to MUSIC practices while also becoming a learning health system to achieve greater success. An LHS approach could further advance MUSIC's goal of being the number one in prostate cancer care in the country.

.

Chapter 7 Conclusion

Learning What Works and What Does Not Work in Prostate Cancer Care

Most of what we have learned in prostate cancer have come from randomized controlled trials, longitudinal cohort studies and from routine clinical practice. Several major randomized control trials and longitudinal cohort studies have been conducted over the past three decades to understand more about the efficacy of prostate cancer screening and treatment. Most of these trials were designed to answer many important questions, including whether screening and treatments for prostate cancer have any mortality benefit. These trials significantly improved our understanding of the disease, while others have added to ongoing uncertainties. Over the past two decades, we have seen considerable progress towards building a greater understanding of prostate cancer – what this disease is and what it is not. However, controversies are widespread in prostate cancer care. Progress is incremental and often contentious as interests, experiences, and training influence experts to arrive at different conclusions despite seeing the same data. The words of Mulley and Barry written over two decades ago continue to reverberate, "*the poorer the evidence, the more discretionary the interpretation, and the more controversial the conclusion*."

The Learning Health System approach offers promise to find clarity in the murky waters of prostate cancer care. Traditional clinical trials tend to evaluate head to head efficacy of one treatment option over another option. Cohort studies observe temporal changes in outcomes but seldom question why a treatment worked on some people and not on others. A Learning Health System offers an opportunity to answer these questions and generate new knowledge by systematically integrating clinical encounter and patient reported outcomes data with external evidence including findings from randomized clinical trials and cohort studies. The vision of converting real-world data through cyclical learning has the potential to continuously inform and improve clinical practice and strengthen the evidence-base, reduce discretionary interpretation of the evidence and support patient decision-making process.

Bibliography

1. SEER Cancer Statistics Reivew 1975-2016. *Lifetime Risk (Percent) of Being Diagnosed with Cancer by Site and Race/Ethnicity*.; 2020. Accessed January 23, 2020. https://surveillance.cancer.gov/devcan/

2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA A Cancer J Clin*. 2020;70(1):7-30. doi:10.3322/caac.21590

3. Brawley OW. Prostate cancer epidemiology in the United States. *World Journal of Urology*. 2012;30(2):195-200. doi:10.1007/s00345-012-0824-2

4. Negoita S, Feuer EJ, Mariotto A, et al. Annual Report to the Nation on the Status of Cancer, part II: Recent changes in prostate cancer trends and disease characteristics: Recent Changes in Prostate Cancer Trends. *Cancer*. 2018;124(13):2801-2814. doi:10.1002/cncr.31549

5. Potosky AL, Feuer EJ, Levin DL. Impact of Screening on Incidence and Mortality of Prostate Cancer in the United States. *Epidemiologic Reviews*. 2001;23(1):181-186. doi:10.1093/oxfordjournals.epirev.a000787

6. Merrill RM, Feuer EJ, Warren JL, Schussler N, Stephenson RA. Role of Transurethral Resection of the Prostate in Population-based Prostate Cancer Incidence Rates. *American Journal of Epidemiology*. 1999;150(8):848-860. doi:10.1093/oxfordjournals.aje.a010090

7. Jemal A, Fedewa SA, Ma J, et al. Prostate Cancer Incidence and PSA Testing Patterns in Relation to USPSTF Screening Recommendations. *JAMA*. 2015;314(19):2054. doi:10.1001/jama.2015.14905

8. Kim EH, Andriole GL. Prostate-specific antigen-based screening: controversy and guidelines. *BMC Med*. 2015;13(1):61. doi:10.1186/s12916-015-0296-5

9. Myers RE, Leader AE, Censits JH, et al. Decision Support and Shared Decision Making About Active Surveillance Versus Active Treatment Among Men Diagnosed with Low-Risk Prostate Cancer: a Pilot Study. *Journal of Cancer Education*. 2018;33(1):180-185. doi:10.1007/s13187-016-1073-7

10. Jahn JL, Giovannucci EL, Stampfer MJ. The high prevalence of undiagnosed prostate cancer at autopsy: implications for epidemiology and treatment of prostate cancer in the Prostate-specific Antigen-era: High prostate cancer prevalence: Research implications in the PSA-ERA. *Int J Cancer*. 2015;137(12):2795-2802. doi:10.1002/ijc.29408

11. Cross DS, Ritter M, Reding DJ. Historical Prostate Cancer Screening and Treatment Outcomes from a Single Institution. *Clinical Medicine & Research*. 2012;10(3):97-105. doi:10.3121/cmr.2011.1042

12. Howlader N, Noone A, Krapcho M, et al. *SEER Cancer Statistics Review, 1975-2016*. National Cancer Institute; 2019. Accessed January 12, 2020. https://seer.cancer.gov/csr/1975_2016/

13. Mariotto AB, Robin Yabroff K, Shao Y, Feuer EJ, Brown ML. Projections of the Cost of Cancer Care in the United States: 2010-2020. *JNCI Journal of the National Cancer Institute*. 2011;103(2):117-128. doi:10.1093/jnci/djq495

14. Hoffman KE, Penson DF, Zhao Z, et al. Patient-Reported Outcomes Through 5 Years for Active Surveillance, Surgery, Brachytherapy, or External Beam Radiation With or Without Androgen Deprivation Therapy for Localized Prostate Cancer. *JAMA*. 2020;323(2):149. doi:10.1001/jama.2019.20675

15. National Institutes of Health. *Estimates of Funding for Various Research, Condition, and Disease Categories (RCDC)*.; 2019. https://report.nih.gov/categorical_spending.aspx

16. Schroeck FR, Jacobs BL, Bhayani SB, Nguyen PL, Penson D, Hu J. Cost of New Technologies in Prostate Cancer Treatment: Systematic Review of Costs and Cost Effectiveness of Robotic-assisted Laparoscopic Prostatectomy, Intensity-modulated Radiotherapy, and Proton Beam Therapy. *European Urology*. 2017;72(5):712-735. doi:10.1016/j.eururo.2017.03.028

17. US Preventive Services Task Force, Grossman DC, Curry SJ, et al. Screening for Prostate Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA*. 2018;319(18):1901. doi:10.1001/jama.2018.3710

18. Etzioni R. Overdiagnosis Due to Prostate-Specific Antigen Screening: Lessons From U.S. Prostate Cancer Incidence Trends. *CancerSpectrum Knowledge Environment*. 2002;94(13):981-990. doi:10.1093/jnci/94.13.981

19. Sturgeon KM, Deng L, Bluethmann SM, et al. A population-based study of cardiovascular disease mortality risk in US cancer patients. *European Heart Journal*. 2019;40(48):3889-3897. doi:10.1093/eurheartj/ehz766

20. Mayor S. Robotic surgery for prostate cancer achieves similar outcomes to open surgery, study shows. *BMJ*. Published online July 27, 2016:i4150. doi:10.1136/bmj.i4150

21. Borza T, Kaufman SR, Shahinian VB, et al. Sharp Decline In Prostate Cancer Treatment Among Men In The General Population, But Not Among Diagnosed Men. *Health Affairs*. 2017;36(1):108-115. doi:10.1377/hlthaff.2016.0739

22. Wilt TJ, Brawer MK, Jones KM, et al. Radical Prostatectomy versus Observation for Localized Prostate Cancer. *New England Journal of Medicine*. 2012;367(3):203-213. doi:10.1056/NEJMoa1113162

23. Mulley AG, Barry MJ. Controversy in managing patients with prostate cancer. *BMJ*. 1998;316(7149):1919-1920. doi:10.1136/bmj.316.7149.1919

24. Cooperberg MR, Broering JM, Carroll PR. Time Trends and Local Variation in Primary Treatment of Localized Prostate Cancer. *Journal of Clinical Oncology*. 2010;28(7):1117-1123. doi:10.1200/JCO.2009.26.0133

25. Institute of Medicine. *Initial National Priorities for Comparative Effectiveness Research*. National Academies Press; 2009. doi:10.17226/12648

26. Houston KA, King J, Li J, Jemal A. Trends in Prostate Cancer Incidence Rates and Prevalence of Prostate Specific Antigen Screening by Socioeconomic Status and Regions in the United States, 2004 to 2013. *Journal of Urology*. 2018;199(3):676-682. doi:10.1016/j.juro.2017.09.103

27. Reese AC, Wessel SR, Fisher SG, Mydlo JH. Evidence of prostate cancer "reverse stage migration" toward more advanced disease at diagnosis: Data from the Pennsylvania Cancer Registry. *Urologic Oncology: Seminars and Original Investigations*. 2016;34(8):335.e21-335.e28. doi:10.1016/j.urolonc.2016.03.014

28. Trogdon JG, Falchook AD, Basak R, Carpenter WR, Chen RC. Total Medicare Costs Associated With Diagnosis and Treatment of Prostate Cancer in Elderly Men. *JAMA Oncol*. 2019;5(1):60. doi:10.1001/jamaoncol.2018.3701

29. Esserman LJ, Thompson IM, Reid B, et al. Addressing overdiagnosis and overtreatment in cancer: a prescription for change. *The Lancet Oncology*. 2014;15(6):e234-e242. doi:10.1016/S1470-2045(13)70598-9

30. Cooperberg MR, Broering JM, Carroll PR. Time Trends and Local Variation in Primary Treatment of Localized Prostate Cancer. *Journal of Clinical Oncology*. 2010;28(7):1117-1123. doi:10.1200/JCO.2009.26.0133

31. Daskivich TJ, Chamie K, Kwan L, et al. Overtreatment of men with low-risk prostate cancer and significant comorbidity. *Cancer*. 2011;117(10):2058-2066. doi:10.1002/cncr.25751

32. PDQ Adult Treatment Editorial Board. Prostate Cancer Treatment (PDQ®): Health Professional Version. In: *PDQ Cancer Information Summaries*. National Cancer Institute (US); 2002. Accessed January 24, 2020. http://www.ncbi.nlm.nih.gov/books/NBK66036/

33. Sartor AO, Fitzpatrick JM. Urologists and oncologists: adapting to a new treatment paradigm in castration-resistant prostate cancer (CRPC): ADAPTING TO A NEW TREATMENT PARADIGM IN CRPC. *BJU International*. 2012;110(3):328-335. doi:10.1111/j.1464-410X.2011.10818.x

34. Hamdy FC, Donovan JL, Lane JA, et al. 10-Year Outcomes after Monitoring, Surgery, or Radiotherapy for Localized Prostate Cancer. *New England Journal of Medicine*. 2016;375(15):1415-1424. doi:10.1056/NEJMoa1606220

35. Carter HB. Prostate-Specific Antigen (PSA) Screening for Prostate Cancer: Revisiting the Evidence. *JAMA*. 2018;319(18):1866. doi:10.1001/jama.2018.4914

36. Penson DF. Quality of Life Outcomes Following Treatment for Localized Prostate Cancer: What's New and What's Not. *European Urology*. 2017;72(6):886-887. doi:10.1016/j.eururo.2017.07.010

37. American Cancer Society. Survival Rates for Prostate Cancer. Published December 18, 2017. https://www.cancer.org/cancer/prostate-cancer/detection-diagnosis-staging/survival-rates.html

38. Aizer AA, Paly JJ, Michaelson MD, et al. Medical Oncology Consultation and Minimization of Overtreatment in Men With Low-Risk Prostate Cancer. *JOP*. 2014;10(2):107-112. doi:10.1200/JOP.2013.000902

39. Hoffman KE, Niu J, Shen Y, et al. Physician Variation in Management of Low-Risk Prostate Cancer: A Population-Based Cohort Study. *JAMA Internal Medicine*. 2014;174(9):1450. doi:10.1001/jamainternmed.2014.3021

40. Tang C, Hoffman KE, Allen PK, et al. Contemporary prostate cancer treatment choices in multidisciplinary clinics referenced to national trends. *Cancer*. 2020;126(3):506-514. doi:10.1002/cncr.32570

41. Fowler, Jr FJ. Comparison of Recommendations by Urologists and Radiation Oncologists for Treatment of Clinically Localized Prostate Cancer. *JAMA*. 2000;283(24):3217. doi:10.1001/jama.283.24.3217

42. Gillitzer R, Hampel C, Thomas C, et al. Bevorzugte Behandlungsoptionen des lokalisierten Prostatakarzinoms von deutschen Urologen und Radioonkologen bei eigener Erkrankung. *Urologe*. 2009;48(4):399-407. doi:10.1007/s00120-008-1928-6

43. Kim SP, Gross CP, Nguyen PY, et al. Specialty bias in treatment recommendations and quality of life among radiation oncologists and urologists for localized prostate cancer. *Prostate Cancer Prostatic Dis*. 2014;17(2):163-169. doi:10.1038/pcan.2014.3

44. Jiang T, Stillson CH, Pollack CE, et al. How Men with Prostate Cancer Choose Specialists: A Qualitative Study. *J Am Board Fam Med*. 2017;30(2):220-229. doi:10.3122/jabfm.2017.02.160163

45. Pollard S, Bansback N, Bryan S. Physician attitudes toward shared decision making: A systematic review. *Patient Education and Counseling*. 2015;98(9):1046-1057. doi:10.1016/j.pec.2015.05.004

46. Elwyn G, Frosch D, Rollnick S. Dual equipoise shared decision making: definitions for decision and behaviour support interventions. *Implementation Sci*. 2009;4(1):75. doi:10.1186/1748-5908-4-75

47. Berry DL, Halpenny B, Wolpin S, et al. Development and Evaluation of the Personal Patient Profile-Prostate (P3P), a Web-Based Decision Support System for Men Newly Diagnosed With Localized Prostate Cancer. *J Med Internet Res*. 2010;12(4):e67. doi:10.2196/jmir.1576

48. Elliott D, Hamdy FC, Leslie TA, et al. Overcoming difficulties with equipoise to enable recruitment to a randomised controlled trial of partial ablation vs radical prostatectomy for unilateral localised prostate cancer. *BJU Int*. 2018;122(6):970-977. doi:10.1111/bju.14432

49. O'Connor AM, Wennberg JE, Legare F, et al. Toward The 'Tipping Point': Decision Aids And Informed Patient Choice. *Health Affairs*. 2007;26(3):716-725. doi:10.1377/hlthaff.26.3.716

50. Wang EH, Gross CP, Tilburt JC, et al. Shared Decision Making and Use of Decision Aids for Localized Prostate Cancer: Perceptions From Radiation Oncologists and Urologists. *JAMA Intern Med*. 2015;175(5):792. doi:10.1001/jamainternmed.2015.63

51. Daum LM, Reamer EN, Ruterbusch JJ, Liu J, Holmes-Rovner M, Xu J. Patient Knowledge and Qualities of Treatment Decisions for Localized Prostate Cancer. *J Am Board Fam Med*. 2017;30(3):288-297. doi:10.3122/jabfm.2017.03.160298

52. Luckenbaugh AN, Miller DC, Ghani KR. Collaborative quality improvement. *Current Opinion in Urology*. 2017;27(4):395-401. doi:10.1097/MOU.0000000000000404

53. Auffenberg GB, Lane BR, Linsell S, et al. A Roadmap for Improving the Management of Favorable Risk Prostate Cancer. *Journal of Urology*. 2017;198(6):1220-1222. doi:10.1016/j.juro.2017.07.085

54. Michigan Urological Surgery Improvement Collaborative. *Roadmap for Management of Men With Favorable-Risk Prostate Cancer*. Michigan Urological Surgery Improvement Collaborative; 2016. Accessed June 2, 2020. https://musicurology.com/wp-content/uploads/2016/12/MUSIC-AS-Roadmap-Patient-Facing_v2.pdf

55. Michigan Urological Surgery Improvement Collaborative. *Introduction to MUSIC's New Performance Measures for Active Surveillance*. Michigan Urological Surgery Improvement Collaborative; 2017. Accessed June 2, 2020. https://musicurology.com/wp-content/uploads/2017/03/MUSIC-Performance-Measurement-Intro_FINALv2.pdf

56. Kaye DR, Qi J, Morgan TM, et al. Association Between Early Confirmatory Testing and the Adoption of Active Surveillance for Men With Favorable-risk Prostate Cancer. *Urology*. 2018;118:127-133. doi:10.1016/j.urology.2018.04.038

57. Hawken SR, Womble PR, Herrel LA, et al. Understanding the Performance of Active Surveillance Selection Criteria in Diverse Urology Practices. *Journal of Urology*. 2015;194(5):1253-1257. doi:10.1016/j.juro.2015.05.014

58. Berry DL. Patient-Reported Symptoms and Quality of Life Integrated into Clinical Cancer Care. *Seminars in Oncology Nursing*. 2011;27(3):203-210. doi:10.1016/j.soncn.2011.04.005

59. Davidson GH, Haukoos JS, Feldman LS. Practical Guide to Assessment of Patient-Reported Outcomes. *JAMA Surg*. Published online January 29, 2020. doi:10.1001/jamasurg.2019.4526

60. Food and Drug Administration (FDA) Center for Drug Education and Research. *Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims*. Food and Drug Administration (FDA); 2009. https://www.fda.gov/media/77832/download

61. Cella DF. *Patient-Reported Outcomes in Performance Measurement*. RTI Press/RTI International; 2015.

62. Snyder CF, Aaronson NK, Choucair AK, et al. Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations. *Qual Life Res*. 2012;21(8):1305-1314. doi:10.1007/s11136-011-0054-x

63. Black N. Patient reported outcome measures could help transform healthcare. *BMJ*. 2013;346(jan28 1):f167-f167. doi:10.1136/bmj.f167

64. Efficace F, Feuerstein M, Fayers P, et al. Patient-reported Outcomes in Randomised Controlled Trials of Prostate Cancer: Methodological Quality and Impact on Clinical Decision Making. *European Urology*. 2014;66(3):416-427. doi:10.1016/j.eururo.2013.10.017

65. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *Journal of Clinical Epidemiology*. 2003;56(5):395-407. doi:10.1016/S0895-4356(03)00044-1

66. Hsu T, Speers CH, Kennecke HF, Cheung WY. The utility of abbreviated patient-reported outcomes for predicting survival in early stage colorectal cancer: Patient-Reported Outcomes and CRC. *Cancer*. 2017;123(10):1839-1847. doi:10.1002/cncr.30511

67. Lardas M, Liew M, van den Bergh RC, et al. Quality of Life Outcomes after Primary Treatment for Clinically Localised Prostate Cancer: A Systematic Review. *European Urology*. 2017;72(6):869-885. doi:10.1016/j.eururo.2017.06.035

68. Squitieri L, Bozic KJ, Pusic AL. The Role of Patient-Reported Outcome Measures in Value-Based Payment Reform. *Value in Health*. 2017;20(6):834-836. doi:10.1016/j.jval.2017.02.003

69. Skolarus TA, Dunn RL, Sanda MG, et al. Minimally Important Difference for the Expanded Prostate Cancer Index Composite Short Form. *Urology*. 2015;85(1):101-106. doi:10.1016/j.urology.2014.08.044

70. Basch E. The Missing Voice of Patients in Drug-Safety Reporting. *N Engl J Med*. 2010;362(10):865-869. doi:10.1056/NEJMp0911494

71. Basch E, Wilfong L, Schrag D. Adding Patient-Reported Outcomes to Medicare's Oncology Value-Based Payment Model. *JAMA*. 2020;323(3):213. doi:10.1001/jama.2019.19970

72. Centers for Medicare & Medicaid Services. Oncology Care First Model: Informal Request for Information. Published November 14, 2019. Accessed March 1, 2020. https://innovation.cms.gov/Files/x/ocf-informalrfi.pdf

73. Locklear T, Weinfurt K, Abernethy A, Flynn K, Riley W. Resource Chapters: Patient-Reported Outcomes. In: Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials. In: *Rethinking Clinical Trials: A Living TextBook of Pragmatic Clinical Trials*. ; 2019. https://rethinkingclinicaltrials.org/resources/patient-reported-outcomes-3/

74. Sohn W, Resnick MJ, Greenfield S, et al. Impact of Adherence to Quality Measures for Localized Prostate Cancer on Patient-reported Health-related Quality of Life Outcomes, Patient Satisfaction, and Treatment-related Complications: *Medical Care*. 2016;54(8):738-744. doi:10.1097/MLR.0000000000000562

75. Gori D, Dulal R, Blayney DW, et al. Utilization of Prostate Cancer Quality Metrics for Research and Quality Improvement: A Structured Review. *The Joint Commission Journal on Quality and Patient Safety*. 2019;45(3):217-226. doi:10.1016/j.jcjq.2018.06.004

76. Burstin H, Leatherman S, Goldmann D. The evolution of healthcare quality measurement in the United States. *J Intern Med*. 2016;279(2):154-159. doi:10.1111/joim.12471

77. Kotronoulas G, Kearney N, Maguire R, et al. What Is the Value of the Routine Use of Patient-Reported Outcome Measures Toward Improvement of Patient Outcomes, Processes of Care, and Health Service Outcomes in Cancer Care? A Systematic Review of Controlled Trials. *JCO*. 2014;32(14):1480-1501. doi:10.1200/JCO.2013.53.5948

78. Sanda MG, Dunn RL, Michalski J, et al. Quality of Life and Satisfaction with Outcome among Prostate-Cancer Survivors. *N Engl J Med*. 2008;358(12):1250-1261. doi:10.1056/NEJMoa074311

79. Donabedian A. Evaluating the quality of medical care. 1966. *Milbank Q*. 2005;83(4):691-729. doi:10.1111/j.1468-0009.2005.00397.x

80. Singh J, Trabulsi EJ, Gomella LG. The Quality-of-Life Impact of Prostate Cancer Treatments. *Curr Urol Rep*. 2010;11(3):139-146. doi:10.1007/s11934-010-0103-y

81. Basch E. The Rationale for Collecting Patient-Reported Symptoms during Routine Chemotherapy. *American Society of Clinical Oncology Educational Book*. 2014;(34):161-165. doi:10.14694/EdBook_AM.2014.34.161

82. Basch E, Deal AM, Kris MG, et al. Symptom Monitoring With Patient-Reported Outcomes During Routine Cancer Treatment: A Randomized Controlled Trial. *JCO*. 2016;34(6):557-565. doi:10.1200/JCO.2015.63.0830

83. Dowrick AS, Wootten AC, Murphy DG, Costello AJ. "We Used a Validated Questionnaire": What Does This Mean and Is It an Accurate Statement in Urologic Research? *Urology*. 2015;85(6):1304-1311. doi:10.1016/j.urology.2015.01.046

84. Litwin MS, Hays RD, Fink A, Ganz PA, Leake B, Brook RH. The UCLA Prostate Cancer Index: Development, Reliability, and Validity of a Health-Related Quality of Life Measure. *Medical Care*. 1998;36(7):1002-1012. doi:10.1097/00005650-199807000-00007

85. Wei JT, Dunn RL, Litwin MS, Sandler HM, Sanda MG. Development and validation of the expanded prostate cancer index composite (EPIC) for comprehensive assessment of health-related quality of life in men with prostate cancer. *Urology*. 2000;56(6):899-905. doi:10.1016/S0090-4295(00)00858-X

86. Orom H, Biddle C, Underwood W, Nelson CJ, Homish DL. What Is a "Good" Treatment Decision? Decisional Control, Knowledge, Treatment Decision Making, and Quality of Life in Men with Clinically Localized Prostate Cancer. *Med Decis Making*. 2016;36(6):714-725. doi:10.1177/0272989X16635633

87. Szymanski KM, Wei JT, Dunn RL, Sanda MG. Development and validation of an abbreviated version of the expanded prostate cancer index composite instrument for measuring health-related quality of life among prostate cancer survivors. *Urology*. 2010;76(5):1245-1250. doi:10.1016/j.urology.2010.01.027

88. Singh K, Tin AL, Dunn RL, Kim T, Vickers AJ. Development and Validation of Crosswalks for Patient-reported Sexual and Urinary Outcomes Between Commonly Used Instruments. *European Urology*. 2019;75(5):723-730. doi:10.1016/j.eururo.2018.12.002

89. Chang P, Szymanski KM, Dunn RL, et al. Expanded Prostate Cancer Index Composite for Clinical Practice: Development and Validation of a Practical Health Related Quality of Life Instrument for Use in the Routine

Clinical Care of Patients With Prostate Cancer. *Journal of Urology*. 2011;186(3):865-872. doi:10.1016/j.juro.2011.04.085

90. Chipman JJ, Sanda MG, Dunn RL, et al. Measuring and Predicting Prostate Cancer Related Quality of Life Changes Using EPIC for Clinical Practice. *Journal of Urology*. 2014;191(3):638-645. doi:10.1016/j.juro.2013.09.040

91. HealthMeasures. Intro to PROMIS®. Published 2020. https://www.healthmeasures.net/explore-measurement-systems/promis/intro-to-promis

92. Cella D, Riley W, Stone A, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*. 2010;63(11):1179-1194. doi:10.1016/j.jclinepi.2010.04.011

93. Thong AE, Ying Poon B, Lee JK, et al. Concordance between patient-reported and physician-reported sexual function after radical prostatectomy. *Urologic Oncology: Seminars and Original Investigations*. 2018;36(2):80.e1-80.e6. doi:10.1016/j.urolonc.2017.09.017

94. Mirza* Mahin, Linsell Susan, Qi Ji, et al. MP66-15 MEASURING THE IMPACT OF A PHYSICIAN-LED COLLABORATIVE ON THE QUALITY OF PROSTATE CANCER CARE: 7 YEARS OF MAKING MUSIC. *Journal of Urology*. 2020;203(Supplement 4):e989-e989. doi:10.1097/JU.0000000000000941.015

95. Tapper* A, Wilson A, Lucas S, et al. MP44-11 PATIENT SELECTION AND OUTCOMES BETWEEN LOW AND HIGH VOLUME SURGEONS IN PERFORMANCE OF RADICAL PROSTATECTOMY IN THE MICHIGAN UROLOGICAL SURGERY IMPROVEMENT COLLABORATIVE (MUSIC). *Journal of Urology*. 2019;201(Supplement 4). doi:10.1097/01.JU.0000556254.87388.a4

96. Michigan Urological Surgery Improvement Collaborative. Michigan Urological Surgery Improvement Collaborative (MUSIC) - Year 8 Progress Report. In: ; 2019. https://musicurology.com/mid-year-report-2019/

97. Agochukwu NQ, Wittmann D, Boileau NR, et al. Validity of the Patient-Reported Outcome Measurement Information System (PROMIS) Sexual Interest and Satisfaction Measures in Men Following Radical Prostatectomy. *JCO*. 2019;37(23):2017-2027. doi:10.1200/JCO.18.01782

98. Chun FKH, Karakiewicz PI, Briganti A, et al. Prostate Cancer Nomograms: An Update. *European Urology*. 2006;50(5):914-926. doi:10.1016/j.eururo.2006.07.042

99. Partin AW, Mangold LA, Lamm DM, Walsh PC, Epstein JI, Pearson JD. Contemporary update of prostate cancer staging nomograms (Partin Tables) for the new millennium. *Urology*. 2001;58(6):843-848. doi:10.1016/S0090-4295(01)01441-8

100. Ross PL, Scardino PT, Kattan MW. A CATALOG OF PROSTATE CANCER NOMOGRAMS. *Journal of Urology*. 2001;165(5):1562-1568. doi:10.1016/S0022-5347(05)66349-5

101. Weaver JK, Kim EH, Vetter JM, et al. Prostate Magnetic Resonance Imaging Provides Limited Incremental Value Over the Memorial Sloan Kettering Cancer Center Preradical Prostatectomy Nomogram. *Urology*. 2018;113:119-128. doi:10.1016/j.urology.2017.10.051

102. Jenkins DA, Martin GP, Sperrin M, et al. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagn Progn Res*. 2021;5(1):1. doi:10.1186/s41512-020-00090-3

103. Vickers AJ, Fearn P, Scardino PT, Kattan MW. Why Can't Nomograms Be More Like Netflix? *Urology*. 2010;75(3):511-513. doi:10.1016/j.urology.2009.07.1265

104. Halabi S, Li C, Luo S. Developing and Validating Risk Assessment Models of Clinical Outcomes in Modern Oncology. *JCO Precision Oncology*. 2019;(3):1-12. doi:10.1200/PO.19.00068

105. Kattan MW, Eastham JA, Stapleton AMF, Wheeler TM, Scardino PT. A Preoperative Nomogram for Disease Recurrence Following Radical Prostatectomy for Prostate Cancer. *JNCI: Journal of the National Cancer Institute*. 1998;90(10):766-771. doi:10.1093/jnci/90.10.766

106. Partin AW, Kattan MW, Subong EN, et al. Combination of prostate-specific antigen, clinical stage, and Gleason score to predict pathological stage of localized prostate cancer. A multi-institutional update. *JAMA*. 1997;277(18):1445-1451.

107. Tosoian JJ, Chappidi M, Feng Z, et al. Prediction of pathological stage based on clinical stage, serum prostate-specific antigen, and biopsy Gleason score: Partin Tables in the contemporary era. *BJU Int*. 2017;119(5):676-683. doi:10.1111/bju.13573

108. Gandaglia G, Fossati N, Dell'Oglio P, Montorsi F, Briganti A. Is there a role for pure clinical prediction models in prostate cancer in the contemporary era? *BJU Int*. 2017;119(5):652-653. doi:10.1111/bju.13833

109. Mehralivand S, Shih JH, Rais-Bahrami S, et al. A Magnetic Resonance Imaging–Based Prediction Model for Prostate Biopsy Risk Stratification. *JAMA Oncol*. 2018;4(5):678. doi:10.1001/jamaoncol.2017.5667

110. Lamberg H, Shankar PR, Singh K, et al. Preoperative Prostate MRI Predictors of Urinary Continence Following Radical Prostatectomy. *Radiology*. Published online January 18, 2022:210500. doi:10.1148/radiol.210500

111. Vickers AJ, Kent M, Scardino PT. Implementation of Dynamically Updated Prediction Models at the Point of Care at a Major Cancer Center: Making Nomograms More Like Netflix. *Urology*. 2017;102:1-3. doi:10.1016/j.urology.2016.10.049

112. Ötleş E, Denton BT, Qu B, et al. Development and Validation of Models to Predict Pathological Outcomes of Radical Prostatectomy in Regional and National Cohorts. *Journal of Urology*. 2022;207(2):358-366. doi:10.1097/JU.0000000000002230

113. Kattan MW, Shariat SF, Andrews B, et al. The Addition of Interleukin-6 Soluble Receptor and Transforming Growth Factor Beta $_1$ Improves a Preoperative Nomogram for Predicting Biochemical Progression in Patients With Clinically Localized Prostate Cancer. *JCO*. 2003;21(19):3573-3579. doi:10.1200/JCO.2003.12.037

114. Kattan MW, Gerds TA. A Framework for the Evaluation of Statistical Prediction Models. *Chest*. 2020;158(1):S29-S38. doi:10.1016/j.chest.2020.03.005

115. Yang Z, Hou Y, Lyu J, Liu D, Chen Z. Dynamic prediction and prognostic analysis of patients with cervical cancer: a landmarking analysis approach. *Annals of Epidemiology*. 2020;44:45-51. doi:10.1016/j.annepidem.2020.01.009

116. Strobl AN, Vickers AJ, Van Calster B, et al. Improving patient prostate cancer risk assessment: Moving from static, globally-applied to dynamic, practice-specific risk calculators. *Journal of Biomedical Informatics*. 2015;56:87-93. doi:10.1016/j.jbi.2015.05.001

117. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*. 2017;24(1):198-208. doi:10.1093/jamia/ocw042

118. Steyerberg EW, Uno H, Ioannidis JPA, et al. Poor performance of clinical prediction models: the harm of commonly applied methods. *Journal of Clinical Epidemiology*. 2018;98:133-143. doi:10.1016/j.jclinepi.2017.11.013

119. Verbakel JY, Steyerberg EW, Uno H, et al. ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. *Journal of Clinical Epidemiology*. 2020;126:207-216. doi:10.1016/j.jclinepi.2020.01.028

120. Boyce S, Fan Y, Watson RW, Murphy TB. Evaluation of prediction models for the staging of prostate cancer. *BMC Med Inform Decis Mak*. 2013;13(1):126. doi:10.1186/1472-6947-13-126

121. Jenkins DA, Sperrin M, Martin GP, Peek N. Dynamic models to predict health outcomes: current status and methodological challenges. *Diagn Progn Res*. 2018;2(1):23. doi:10.1186/s41512-018-0045-2

122. Finlayson SG, Subbaswamy A, Singh K, et al. The Clinician and Dataset Shift in Artificial Intelligence. *N Engl J Med*. 2021;385(3):283-286. doi:10.1056/NEJMc2104626

123. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*. Published online November 19, 2019:kxz041. doi:10.1093/biostatistics/kxz041

124. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. *Pattern Recognition*. 2012;45(1):521-530. doi:10.1016/j.patcog.2011.06.019

125. Quinonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND. *Dataset Shift in Machine Learning*.; 2008.

126. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, On behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230. doi:10.1186/s12916-019-1466-7

127. Friedman CP, Wong AK, Blumenthal D. Achieving a Nationwide Learning Health System. *Science Translational Medicine*. 2010;2(57):57cm29-57cm29. doi:10.1126/scitranslmed.3001456

128. Friedman C, Rubin J, Brown J, et al. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. *Journal of the American Medical Informatics Association*. Published online October 23, 2014:amiajnl-2014-002977. doi:10.1136/amiajnl-2014-002977

129. Hamdy FC, Donovan JL, Lane JA, et al. 10-Year Outcomes after Monitoring, Surgery, or Radiotherapy for Localized Prostate Cancer. *N Engl J Med*. 2016;375(15):1415-1424. doi:10.1056/NEJMoa1606220

130. Wilt TJ, Jones KM, Barry MJ, et al. Follow-up of Prostatectomy versus Observation for Early Prostate Cancer. *N Engl J Med*. 2017;377(2):132-142. doi:10.1056/NEJMoa1615869

131. Wilt TJ, Vo TN, Langsetmo L, et al. Radical Prostatectomy or Observation for Clinically Localized Prostate Cancer: Extended Follow-up of the Prostate Cancer Intervention Versus Observation Trial (PIVOT). *European Urology*. 2020;77(6):713-724. doi:10.1016/j.eururo.2020.02.009

132. Barocas DA, Alvarez J, Resnick MJ, et al. Association Between Radiation Therapy, Surgery, or Observation for Localized Prostate Cancer and Patient-Reported Outcomes After 3 Years. *JAMA*. 2017;317(11):1126. doi:10.1001/jama.2017.1704

133. Donovan JL, Hamdy FC, Lane JA, et al. Patient-Reported Outcomes after Monitoring, Surgery, or Radiotherapy for Prostate Cancer. *N Engl J Med*. 2016;375(15):1425-1437. doi:10.1056/NEJMoa1606221

134. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *Journal of the American Medical Informatics Association*. 2017;24(6):1052-1061. doi:10.1093/jamia/ocx030

135. Auffenberg et al. - 2021 - Evaluation of Patient- and Surgeon-Specific Variat.pdf.

136. Brown VA. An Introduction to Linear Mixed-Effects Modeling in R. *Advances in Methods and Practices in Psychological Science*. 2021;4(1):251524592096035. doi:10.1177/2515245920960351

137. Tasca GA, Illing V, Joyce AS, Ogrodniczuk JS. Three-level multilevel growth models for nested change data: A guide for group treatment researchers. *Psychotherapy Research*. 2009;19(4-5):453-461. doi:10.1080/10503300902933188

138. Womble PR, Montie JE, Ye Z, Linsell SM, Lane BR, Miller DC. Contemporary Use of Initial Active Surveillance Among Men in Michigan with Low-risk Prostate Cancer. *European Urology*. 2015;67(1):44-50. doi:10.1016/j.eururo.2014.08.024

139. Auffenberg GB, Linsell S, Dhir A, et al. Comparison of Pathological Outcomes for Men with Low Risk Prostate Cancer from Diverse Practice Settings: Similar Results from Immediate Prostatectomy or Initial Surveillance with Delayed Prostatectomy. *Journal of Urology*. 2016;196(5):1415-1421. doi:10.1016/j.juro.2016.05.095

140. Liu G, Andreev VP, Helmuth ME, et al. Symptom Based Clustering of Men in the LURN Observational Cohort Study. *Journal of Urology*. 2019;202(6):1230-1239. doi:10.1097/JU.0000000000000354

141. Andreev VP, Liu G, Yang CC, et al. Symptom Based Clustering of Women in the LURN Observational Cohort Study. *Journal of Urology*. 2018;200(6):1323-1331. doi:10.1016/j.juro.2018.06.068

142. Auffenberg GB, Ghani KR, Ramani S, et al. askMUSIC: Leveraging a Clinical Registry to Develop a New Machine Learning Model to Inform Patients of Prostate Cancer Treatments Chosen by Similar Men. *European Urology*. 2019;75(6):901-907. doi:10.1016/j.eururo.2018.09.050

143. Agochukwu-Mmonu N, Murali A, Wittmann D, et al. Development and Validation of Dynamic Multivariate Prediction Models of Sexual Function Recovery in Patients with Prostate Cancer Undergoing Radical Prostatectomy: Results from the MUSIC Statewide Collaborative. *European Urology Open Science*. 2022;40:1-8. doi:10.1016/j.euros.2022.03.009

144. Davis SE, Greevy RA, Lasko TA, Walsh CG, Matheny ME. Detection of calibration drift in clinical prediction models to inform model updating. *Journal of Biomedical Informatics*. 2020;112:103611. doi:10.1016/j.jbi.2020.103611

145. Vickers AJ. Prediction models in cancer care. *CA: A Cancer Journal for Clinicians*. Published online 2011:n/a-n/a. doi:10.3322/caac.20118

146. Van Calster B, Vickers AJ. Calibration of Risk Prediction Models: Impact on Decision-Analytic Performance. *Med Decis Making*. 2015;35(2):162-169. doi:10.1177/0272989X14547233

147. Lindhiem O, Petersen IT, Mentch LK, Youngstrom EA. The Importance of Calibration in Clinical Psychology. *Assessment*. 2020;27(4):840-854. doi:10.1177/1073191117752055

148. Vickers AJ, Cronin AM. Everything You Always Wanted to Know About Evaluating Prediction Models (But Were Too Afraid to Ask). *Urology*. 2010;76(6):1298-1301. doi:10.1016/j.urology.2010.06.019

149. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer International Publishing; 2019. doi:10.1007/978-3-030-16399-0

150. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology*. 2016;74:167-176. doi:10.1016/j.jclinepi.2015.12.005

151. National Cancer Institute. Joinpoint Trend Analysis Software. Published April 21, 2020. Accessed February 23, 2021. https://surveillance.cancer.gov/joinpoint/

152. Kim HJ, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications to cancer rates. *Stat Med*. 2000;19(3):335-351. doi:10.1002/(sici)1097-0258(20000215)19:3<335::aid-sim336>3.0.co;2-z

153. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*. 2020;27(4):621-633. doi:10.1093/jamia/ocz228

154. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Statist Med*. 2014;33(3):517-535. doi:10.1002/sim.5941

155. Stevens RJ, Poppe KK. Validation of clinical prediction models: what does the "calibration slope" really measure? *Journal of Clinical Epidemiology*. 2020;118:93-99. doi:10.1016/j.jclinepi.2019.09.016

156. Davis SE, Greevy RA, Lasko TA, Walsh CG, Matheny ME. Comparison of Prediction Model Performance Updating Protocols: Using a Data-Driven Testing Procedure to Guide Updating. *AMIA Annu Symp Proc*. 2019;2019:1002-1010.

157. Walsh CG, Sharman K, Hripcsak G. Beyond discrimination: A comparison of calibration methods and clinical usefulness of predictive models of readmission risk. *Journal of Biomedical Informatics*. 2017;76:9-18. doi:10.1016/j.jbi.2017.10.008

158. Nyarangi-Dix JN, Radtke JP, Hadaschik B, Pahernik S, Hohenfellner M. Impact of Complete Bladder Neck Preservation on Urinary Continence, Quality of Life and Surgical Margins After Radical Prostatectomy: A Randomized, Controlled, Single Blind Trial. *Journal of Urology*. 2013;189(3):891-898. doi:10.1016/j.juro.2012.09.082

159. Porpiglia F, Bertolo R, Manfredi M, et al. Total Anatomical Reconstruction During Robot-assisted Radical Prostatectomy: Implications on Early Recovery of Urinary Continence. *European Urology*. 2016;69(3):485-495. doi:10.1016/j.eururo.2015.08.005

160. Michl U, Tennstedt P, Feldmeier L, et al. Nerve-sparing Surgery Technique, Not the Preservation of the Neurovascular Bundles, Leads to Improved Long-term Continence Rates After Radical Prostatectomy. *European Urology*. 2016;69(4):584-589. doi:10.1016/j.eururo.2015.07.037

161. Lin D, O'Callaghan ME, David R, et al. *Does Urethral Length Affect Continence Outcomes Following Robot Assisted Laparoscopic Radical Prostatectomy (RALP)?* In Review; 2019. doi:10.21203/rs.2.17860/v1

162. Trieu D, Ju IE, Chang SB, Mungovan SF, Patel MI. Surgeon case volume and continence recovery following radical prostatectomy: a systematic review. *ANZ Journal of Surgery*. 2021;91(4):521-529. doi:10.1111/ans.16491

163. Clements MB, Gmelich CC, Vertosick EA, et al. Have urinary function outcomes after radical prostatectomy improved over the past decade? *Cancer*. 2022;128(5):1066-1073. doi:10.1002/cncr.33994

164. Carlsson S, Berglund A, Sjoberg D, et al. Effects of surgeon variability on oncologic and functional outcomes in a population-based setting. *BMC Urol*. 2014;14(1):25. doi:10.1186/1471-2490-14-25

165. Vickers A, Savage C, Bianco F, et al. Cancer Control and Functional Outcomes After Radical Prostatectomy as Markers of Surgical Quality: Analysis of Heterogeneity Between Surgeons at a Single Cancer Center. *European Urology*. 2011;59(3):317-322. doi:10.1016/j.eururo.2010.10.045

166. Wei JT, Dunn RL, Marcovich R, Montie JE, Sanda MG. PROSPECTIVE ASSESSMENT OF PATIENT REPORTED URINARY CONTINENCE AFTER RADICAL PROSTATECTOMY. *Journal of Urology*. 2000;164(3 Part 1):744-748. doi:10.1016/S0022-5347(05)67294-1

167. Nyberg M, Sjoberg DD, Carlsson SV, et al. Surgeon heterogeneity significantly affects functional and oncological outcomes after radical prostatectomy in the Swedish LAPPRO trial. *BJU International*. 2021;127(3):361-368. doi:10.1111/bju.15238

168. Auffenberg GB, Qi J, Dunn RL, et al. Evaluation of Patient- and Surgeon-Specific Variations in Patient-Reported Urinary Outcomes 3 Months After Radical Prostatectomy From a Statewide Improvement Collaborative. *JAMA Surg*. 2021;156(3):e206359. doi:10.1001/jamasurg.2020.6359

169. Bowen ME, Neuhauser D. Understanding and managing variation: three different perspectives. *Implementation Sci*. 2013;8(S1):S1. doi:10.1186/1748-5908-8-S1-S1

170. Fossati N, Di Trapani E, Gandaglia G, et al. Assessing the Impact of Surgeon Experience on Urinary Continence Recovery After Robot-Assisted Radical Prostatectomy: Results of Four High-Volume Surgeons. *Journal of Endourology*. 2017;31(9):872-877. doi:10.1089/end.2017.0085

171. Trinh QD, Bjartell A, Freedland SJ, et al. A Systematic Review of the Volume–Outcome Relationship for Radical Prostatectomy. *European Urology*. 2013;64(5):786-798. doi:10.1016/j.eururo.2013.04.012

172. Patel VR, Coelho RF, Chauhan S, et al. Continence, potency and oncological outcomes after robotic-assisted radical prostatectomy: early trifecta results of a high-volume surgeon: TRIFECTA OUTCOMES AFTER RARP. *BJU International*. 2010;106(5):696-702. doi:10.1111/j.1464-410X.2010.09541.x

173. Steinsvik EAS, Axcrona K, Angelsen A, et al. Does a surgeon's annual radical prostatectomy volume predict the risk of positive surgical margins and urinary incontinence at one-year follow-up? - Findings from a prospective national study. *Scandinavian Journal of Urology*. 2013;47(2):92-100. doi:10.3109/00365599.2012.707684

174. Kim B, Merchant M, Slezak J, et al. MP15-12 THE IMPACT OF SURGICAL CASELOAD VOLUME ON QUALITY OF LIFE IN MEN AFTER ROBOT-ASSISTED RADICAL PROSTATECTOMY. *Journal of Urology*. 2014;191(4S). doi:10.1016/j.juro.2014.02.565

175. Van den Broeck T, Oprea-Lager D, Moris L, et al. A Systematic Review of the Impact of Surgeon and Hospital Caseload Volume on Oncological and Nononcological Outcomes After Radical Prostatectomy for Nonmetastatic Prostate Cancer. *European Urology*. 2021;80(5):531-545. doi:10.1016/j.eururo.2021.04.028

176. Bianco FJ, Riedel ER, Begg CB, Kattan MW, Scardino PT. VARIATIONS AMONG HIGH VOLUME SURGEONS IN THE RATE OF COMPLICATIONS AFTER RADICAL PROSTATECTOMY: FURTHER EVIDENCE THAT TECHNIQUE MATTERS. *Journal of Urology*. 2005;173(6):2099-2103. doi:10.1097/01.ju.0000158163.21079.66

177. Barocas DA, Mitchell R, Chang SS, Cookson MS. Impact of surgeon and hospital volume on outcomes of radical prostatectomy. *Urologic Oncology: Seminars and Original Investigations*. 2010;28(3):243-250. doi:10.1016/j.urolonc.2009.03.001

178. Wilt TJ, Shamliyan TA, Taylor BC, MacDonald R, Kane RL. Association Between Hospital and Surgeon Radical Prostatectomy Volume and Patient Outcomes: A Systematic Review. *Journal of Urology*. 2008;180(3):820-829. doi:10.1016/j.juro.2008.05.010

179. Yang L, Lee JA, Heer E, et al. One-year urinary and sexual outcome trajectories among prostate cancer patients treated by radical prostatectomy: a prospective study. *BMC Urol*. 2021;21(1):81. doi:10.1186/s12894-021-00845-0

180. Goldenberg MG, Goldenberg L, Grantcharov TP. Surgeon Performance Predicts Early Continence After Robot-Assisted Radical Prostatectomy. *Journal of Endourology*. 2017;31(9):858-863. doi:10.1089/end.2017.0284

181. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models using lme4. *arXiv:14065823 [stat]*. Published online June 23, 2014. Accessed February 7, 2022. http://arxiv.org/abs/1406.5823

182. Detry MA, Ma Y. Analyzing Repeated Measurements Using Mixed Models. *JAMA*. 2016;315(4):407. doi:10.1001/jama.2015.19394

183. Raudenbush SW, Bryk AS. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Vol 1. sage; 2002.

184. Brauer M, Curtin JJ. Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*. 2018;23(3):389-411. doi:10.1037/met0000159

185. Walsh PC, Marschke P, Ricker D, Burnett AL. Patient-reported urinary continence and sexual function after anatomic radical prostatectomy. *Urology*. 2000;55(1):58-61. doi:10.1016/S0090-4295(99)00397-0

186. Hu JC, Elkin EP, Pasta DJ, et al. Predicting Quality of Life After Radical Prostatectomy: Results From CaPSURE. *Journal of Urology*. 2004;171(2):703-708. doi:10.1097/01.ju.0000107964.61300.f6

187. Nyberg M, Sjoberg DD, Carlsson SV, et al. Surgeon heterogeneity significantly affects functional and oncological outcomes after radical prostatectomy in the Swedish LAPPRO trial. *BJU International*. 2021;127(3):361-368. doi:10.1111/bju.15238

188. Begg CB, Riedel ER, Bach PB, et al. Variations in Morbidity after Radical Prostatectomy. *N Engl J Med*. 2002;346(15):1138-1144. doi:10.1056/NEJMsa011788

189. Breiman L. Random Forests. *Machine Learning*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

190. Breiman L. Bagging predictors. *Mach Learn*. 1996;24(2):123-140. doi:10.1007/BF00058655

191. Ishwaran H. The effect of splitting on random forests. *Mach Learn*. 2015;99(1):75-118. doi:10.1007/s10994-014-5451-2

192. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. In: *An Introduction to Statistical Learning*. Springer Texts in Statistics. Springer US; 2021:1-14. doi:10.1007/978-1-0716-1418-1_1

193. Liss MA, Osann K, Canvasser N, et al. Continence Definition After Radical Prostatectomy Using Urinary Quality of Life: Evaluation of Patient Reported Validated Questionnaires. *Journal of Urology*. 2010;183(4):1464-1468. doi:10.1016/j.juro.2009.12.009

194. Vickers AJ, Kent M, Mulhall J, Sandhu J. Counseling the Post–radical Prostatectomy Patients About Functional Recovery: High Predictiveness of Current Status. *Urology*. 2014;84(1):158-163. doi:10.1016/j.urology.2014.02.049

195. Hurwitz LM, Cullen J, Kim DJ, et al. Longitudinal regret after treatment for low- and intermediate-risk prostate cancer: Regret After Prostate Cancer Treatment. *Cancer*. 2017;123(21):4252-4258. doi:10.1002/cncr.30841

196. Wallis CJD, Zhao Z, Huang LC, et al. Association of Treatment Modality, Functional Outcomes, and Baseline Characteristics With Treatment-Related Regret Among Men With Localized Prostate Cancer. *JAMA Oncol*. 2022;8(1):50. doi:10.1001/jamaoncol.2021.5160

197. Subbaswamy A, Saria S. Counterfactual Normalization: Proactively Addressing Dataset Shift Using Causal Mechanisms. In: *UAI*. ; 2018:947-957.

198. Otles E, Oh J, Li B, et al. Mind the performance gap: examining dataset shift during prospective validation. In: *Machine Learning for Healthcare Conference*. PMLR; 2021:506-534.

199. Cooperberg MR, Carroll PR. Trends in Management for Patients With Localized Prostate Cancer, 1990-2013. *JAMA*. 2015;314(1):80. doi:10.1001/jama.2015.6036

200. Fletcher SA, von Landenberg N, Cole AP, et al. Contemporary national trends in prostate cancer risk profile at diagnosis. *Prostate Cancer Prostatic Dis*. 2020;23(1):81-87. doi:10.1038/s41391-019-0157-y

201. Liu Y, Hall IJ, Filson C, Howard DH. Trends in the use of active surveillance and treatments in Medicare beneficiaries diagnosed with localized prostate cancer. *Urologic Oncology: Seminars and Original Investigations*. 2021;39(7):432.e1-432.e10. doi:10.1016/j.urolonc.2020.11.024

202. Vince RA, Sun Y, Mahal B, et al. The Impact of a Statewide Active Surveillance Initiative: A Roadmap for Increasing Active Surveillance Utilization Nationwide. *European Urology*. Published online June 2022:S0302283822024058. doi:10.1016/j.eururo.2022.05.028

203. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer International Publishing; 2019. doi:10.1007/978-3-030-16399-0

204. Hickey GL, Grant SW, Caiado C, et al. Dynamic Prediction Modeling Approaches for Cardiac Surgery. *Circ: Cardiovascular Quality and Outcomes*. 2013;6(6):649-658. doi:10.1161/CIRCOUTCOMES.111.000012

205. Mohler JL, Antonarakis ES, Armstrong AJ, et al. Prostate Cancer, Version 2.2019, NCCN Clinical Practice Guidelines in Oncology. *Journal of the National Comprehensive Cancer Network*. 2019;17(5):479-505. doi:10.6004/jnccn.2019.0023

206. Moons KGM, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart*. 2012;98(9):683-690. doi:10.1136/heartjnl-2011-301246

207. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014;35(29):1925-1931. doi:10.1093/eurheartj/ehu207

208. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845.

209. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12(1):77. doi:10.1186/1471-2105-12-77

210. Riley RD, Ensor J, Snell KIE, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ*. Published online June 22, 2016:i3140. doi:10.1136/bmj.i3140

211. Lenert MC, Matheny ME, Walsh CG. Prognostic models will be victims of their own success, unless…. *Journal of the American Medical Informatics Association*. 2019;26(12):1645-1650. doi:10.1093/jamia/ocz145

212. Institute of Medicine. *The Learning Healthcare System: Workshop Summary (IOM Roundtable on Evidence-Based Medicine)*. National Academies Press; 2007:11903. doi:10.17226/11903

213. Agency of Healthcare Research and Quality. Learning Health Systems. Published November 2017. https://www.ahrq.gov/professionals/systems/learning-health-systems/index.html

214. Nwaru BI, Friedman C, Halamka J, Sheikh A. Can learning health systems help organisations deliver personalised care? *BMC Med*. 2017;15(1):177. doi:10.1186/s12916-017-0935-0

215. Michigan Urological Surgery Improvement Collaborative (MUSIC). Program Summary. Published December 2014. https://musicurology.com/wp-content/uploads/2014/12/MUSIC-Catalog-2015.pdf

216. Montie JE, Linsell SM, Miller DC. Quality of Care in Urology and the Michigan Urological Surgery Improvement Collaborative. *Urology Practice*. 2014;1(2):74-78. doi:10.1016/j.urpr.2014.04.003

217. Miller DC, Murtagh DS, Suh RS, Knapp PM, Dunn RL, Montie JE. Establishment of a Urological Surgery Quality Collaborative. *Journal of Urology*. 2010;184(6):2485-2490. doi:10.1016/j.juro.2010.08.015

218. Ghani KR, Miller DC. Variation in Prostate Cancer Care. *JAMA*. 2015;313(20):2066. doi:10.1001/jama.2015.0607

219. Morain SR, Kass NE, Grossmann C. What allows a health care system to become a learning health care system: Results from interviews with health system leaders: What allows a health care system to become a learning health care system: results from interviews with health system leaders. *Learn Health Sys*. 2017;1(1):e10015. doi:10.1002/lrh2.10015