Comparing Traditional Modeling Approaches Versus Predictive Analytics Methods for Predicting Multiple Sclerosis Relapse and All-Cause Urgent Care

Submitted to the Faculty of the
College of Health Sciences
University of Indianapolis

In partial fulfillment of the requirements for the degree
Doctor of Health Science
By: Karen Walsh, MS, MBA

Approved by:

Elizabeth Moore, PhD _____
Committee Chair

Brant J. Oliver, PhD, MS, MPH, FNP-BC, PMHN-BC _____
Committee Member

Johnathon Kyle Armstrong, PhD _____
Committee Member

Accepted by:

Laura Santurri, PhD, MPH, CPH _____
Director, DHSc Program
Chair, Interprofessional Health & Aging Studies
University of Indianapolis

Stephanie Kelly, PT, PhD _____
Dean, College of Health Sciences
University of Indianapolis

Comparing Traditional Modeling Approaches versus Predictive Analytics Methods for Predicting Multiple Sclerosis Relapse and All-Cause Urgent Care

Karen Walsh

University of Indianapolis

Abstract

Multiple sclerosis is a complex and costly chronic ("3C") condition that currently has no cure. In a condition like multiple sclerosis, which has an unpredictable course, the use of predictive analytics could help health systems learn better, faster, and to improve more effectively and predict rather than react to emerging health needs for people with MS. This study compared traditional statistical methods to different predictive analytics methods on two separate endpoints, MS relapse and all-cause urgent care. Binary logistic regression was compared with other machine learning models, specifically ridge, least absolute shrinkage and selection operator (LASSO), and random forest. Results indicated when comparing relapse indices across models' random forest significantly outperformed logistic regression and other machine learning algorithms ($\Delta perf_A$ =27.1%, $\Delta perf_M$ =27.5%). However, for $\Delta perf_{F,}$, logistic regression and random forest performed relatively the same. Ridge and LASSO outperformed logistic regression ($\Delta perf_{M1}$ =0.9%, $\Delta perf_{M2}$ =9.4%, $\Delta perf_{F2}$=25.8%) respectively. Results indicated when comparing all-cause urgent care indices across models, logistic regression performed similarly to random forest and LASSO ($\Delta perf_A$ = -1.1%, $\Delta perf_M$ =1.58%, $\Delta perf_{A1}$ = -0.5%). Ridge performed worse overall compared to logistic regression ($\Delta perf_{A2}$ = -17.8%, $\Delta perf_{M2}$ = -3.84%).

Acknowledgements

My doctoral journey is non-traditional, and began in 2014 at Thomas Jefferson University as a PhD student and wrapped up seven years later at University of Indianapolis. This long journey was due to family priorities and a big corporate job which made it almost impossible to move forward as a doctoral student. Thomas Jefferson University, College of Population Health gave me a job opportunity as a faculty member in 2018, to continue to move forward and complete my doctorate.

First and foremost, I am extraordinary grateful to my husband, Thomas Walsh, my biggest supporter and fan throughout this time. Not only did he support me during my doctoral journey, but a second master's degree. His willingness to help with family and household obligations, when I was too busy studying, writing, or doing homework was invaluable. He was also a shoulder to cry on when I was overwhelmed by all of my responsibilities, and always found a way to make me laugh. You are the best.

I would also like to thank my chair, Dr. Moore, who has been supportive during this process and has made me appreciate APA formatting and become a better academic writer. I want to also thank Dr. Oliver, not only as a committee member, but as a mentor. I started working with Dr. Oliver in 2018 with MS-CQI and he has been instrumental on guiding and supporting me in my academic growth and an extraordinary research experience. My final acknowledgement goes to Dr. Kyle Armstrong, who pushed me to the next methodological level from a data science perspective during this journey. I am grateful for the time he spent with me to better understand complex data science paradigms and R programming language.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Multiple sclerosis (MS) is one of the most common chronic neurological conditions for adults, with a prevalence of nearly one million people in the United States (Wallin et al., 2019). While the exact pathologic mechanism for MS remains incompletely understood, it involves an immune-mediated process in which an abnormal response of the body's immune system is directed against the central nervous system (CNS; National Multiple Sclerosis Society, 2019a). MS is a complex and costly chronic ("3C") condition that currently has no cure. The disease course is often debilitating, disabling, and unpredictable and causes functional and symptomatic impairments including sensory, motor, cognitive, and psychiatric problems and debilitating fatigue (Bozkaya, Livingston, Migliaccio-Walle, & Odom, 2017).

There are four different types of MS, relapsing-remitting (RRMS), primary progressive (PPMS), secondary progressive (SPMS), and progressive relapsing (PRMS; National Multiple Sclerosis Society, 2019b). Relapsing-remitting MS is the most common disease course with 85% of people initially diagnosed with RRMS (National Multiple Sclerosis Society, 2019b). This type of MS is characterized by relapses of new or increasing neurologic symptoms which are followed by periods of complete or partial recovery (National Multiple Sclerosis Society, 2019b). Primary progressive MS is defined as worsening neurologic function from the onset of symptoms without early relapses or remissions. About 15% of people with MS have PPMS (National Multiple Sclerosis Society, 2019b). Secondary progressive MS follows an initial relapsing-remitting disease course. People that have RRMS may at some point transition to SPMS where there is a progressive worsening of neurologic function over time (National Multiple Sclerosis Society, 2019b).

MS causes a socioeconomic burden to the individual and society due to the disability of the disease (O'Connell et al., 2014). The average total cost of care over 12 months for individuals with MS was $51,692 for non-relapsing MS, while the cost for relapse-remitting MS was $58,648 (Jones, Pike, Marshall, & Ye, 2016) The most common form of this disease is relapsing MS with a relapse of approximately one per year (Vollmer, 2007). Research shows that participants with relapsing MS had significantly higher annual costs for physician consultations ($464; $p < .05$), hospitalizations ($3,693; $p < .05$) and total MS costs ($4,390; $p < .05$) when compared to participants with non-relapsing MS (Iezzoni & Ngo, 2007; Jones et al., 2016).

Clinical Quality Improvement

The proposed study is part of an MS clinical quality improvement research collaborative, the Multiple Sclerosis Continuous Quality Improvement (MS-CQI) Collaborative, which is the parent study (See Appendix A). MS-CQI is an improvement research collaborative that employs a combination of an improvement collaborative structure and a step-wedge randomized design to test the comparative effectiveness of quality improvement (QI) intervention versus usual care. The Health Foundation (2009) describes a QI collaborative as requiring five critical features: (1) there is a specified topic; (2) clinical experts and experts in quality improvement provide ideas and support for improvement; (3) there are multi-professional teams from multiple sites; (4) there is a model for improvement that focuses on setting clear and measurable targets; (5) the collaborative process involves a series of structured activities (de Silva, 2014; Hulscher, Schouten, & Grol, 2009).

Clinical QI collaboratives can facilitate systems-level change in real-world settings (de Silva, 2014; Hulscher et al., 2009.), but data from these collaboratives need to be systematically vetted and analyzed using rigorous research designs. MS-CQI is implementing descriptive

statistics as well as inferential outcomes analysis across four centers to compare performance. Statistical process control (SPC) benchmarking with variation analysis by center is also being conducted. Predictive analytics and more advanced statistical techniques can complement and may improve upon this current standard.

Predictive Analytics

The proposed research study expanded upon the learning health systems (LHS) model even further by moving this work towards predictive analytics. MS-CQI was created as the first LHS improvement research collaborative for MS, and is addressing some of these deficiencies by utilizing a learning health system model that employs integrated improvement and research methods, including individual and population levels of analysis and a randomized comparative effectiveness design (Oliver, 2019). MS-CQI represents the vanguard of a shift towards the inclusion of systems-level approaches to improve MS care and demonstrate value.

The Institute of Medicine (IOM) and the Institute for Healthcare Improvement (IHI) have called for a new systems-oriented focus and continuous improvement culture in the United States (IOM, 2001; IHI, n.d.). Nelson's Balanced Measures "Clinical Value Compass" framework is commonly used in healthcare QI because it can specify process, measure systems-level quality, and value (Lindblad et al., 2017; Nelson et al., 1995). The framework has four categories: (a) clinical outcomes; (b) functional health; (c) patient experience and satisfaction; and (d) utilization (Nelson et al., 2016). Predictive analytics could optimize the feed-forward and feedback aspects of the MS LHS. In a condition like MS, which has an unpredictable course, the use of predictive analytics could help health systems learn better, faster, and to improve more effectively and predict rather than react to emerging health needs for people with MS.

Predictive analytics introduces a new culture in data analysis. The evolution of algorithmic modeling has grown tremendously outside of the field of statistics (Breiman, 2001). These algorithms can be used on various datasets and be more accurate than existing data modeling approaches (Breiman, 2001). Model selection is different in predictive analytics versus traditional statistical modeling approaches. Traditional statistical methods allow one to determine, based on subject matter knowledge and/or simple descriptive or inferential statistics, what predictors to include in a model. Some of these methods include linear regression and logistic regression (James, Witten, Hastie, & Tibshirani, 2013). In predictive analytics the combination of test/training data splits or cross-validation, one can assess model error reliably (James et al., 2013). Predictive analytics methods, including regularization, and ensemble approaches can deal with far larger numbers of predictors in an automated fashion (James et al., 2013). Some of these predictive analytics methods include ridge regression and least absolute shrinkage and selection operator (LASSO) regression (James et al., 2013). Ridge regression and LASSO are both shrinkage methods and can be used with many features in the model without overfitting (James et al., 2013). Overfitting is when the model is trained and fits well on the training dataset, but performs poorly on the test dataset and becomes less generalizable. Both of these models shrink or regularize the coefficient estimates or shrinks them towards zero (James et al., 2013). By shrinking coefficient estimates, this improves the fit of the model by reducing the model's variance (James et al., 2013).

Problem Statement

Historically, MS care has been studied and improved through basic and clinical trials research and epidemiological studies on population outcomes and quality of life. MS registries such as NARCOMS, MS PATH, and MS link have been developed. Only very recently has MS

entered the realm of improvement science and population health. The MS-CQI is the first multi-center exemplar using a learning health system model that employs integrated improvement and research methods, including individual and population levels of analysis (Oliver, 2019). In comparison, the MS-Advance study was a single center, non-randomized study and used a patient centered medical home as their intervention (Meninno et al., 2018). A multi-center approach is needed for a successful collaborative model.

There is a dearth of research on predicting relapse in MS QI collaboratives using predictive analytics, as well as predicting MS all-cause UC utilization. This study is significant because it analyzes outcomes of a multicenter MS quality improvement collaborative as well as comparing traditional statistical methods to predictive analytics for predicting MS relapse and all-cause urgent care (UC) utilization. Relapse and UC utilization are aspects of MS that are costly and disabling, and to date, have been largely unpredictable. For this study all-cause, UC utilization included all-cause emergency department (ED) visits and all-cause UC visits.

Purpose Statement

This study focused on predictive modeling of MS relapse and all-cause UC utilization by comparing standard statistical and predictive analytics approaches. The first research aim is to determine whether results are similar between traditional statistical methods and predictive analytics methods when predicting MS relapse. The second research aim is to determine whether traditional statistical methods versus predictive analytics methods results are similar when predicting MS all cause-UC utilization.

Literature Review

This study analyzes outcomes of a multicenter MS quality improvement collaborative by comparing traditional statistical methods to predictive analytics for predicting MS relapse rate

and all-cause UC. The literature review summarizes available evidence on: (a) quality improvement collaboratives; (b) MS predictive modeling on relapse and all-cause UC utilization; and (c) studies on predictive model comparisons.

Relapsing-remitting MS. Relapse-remitting MS accounts for approximately 85% of MS cases (Vollmer, 2017). RRMS is characterized by periods of exacerbations followed by partial or full recovery to prior baseline. Longitudinally, a subset of people with RRMS progress into a progressive stage called SPMS in which relapses cease, neurological progression continues, and recovery is rare (Vollmer, 2007). MS relapses occur on average about 1.1 per year early in the course of the disease; however, this rate appears to decrease as the disease progresses (Vollmer, 2007). Immunotherapy disease-modifying treatments (DMTs) can dramatically reduce the relapse rate as well as related brain magnetic resonance imaging (MRI) changes and functional neurological progression of disability (Stoppe Busch, Krizek, & Then Bergh, 2017); however, DMTs do not cure MS (Stoppe et al., 2017). Disease-modifying therapies are not used to control MS symptoms.

In the early stages of MS, relapses are more likely to resolve to prior baseline. However, as the disease progresses, relapses result in incomplete recovery and progressive disability and can have a significant impact on the individual's quality of life (O'Connell et al., 2014). Relapses can result in economic cost of health services such as ED visits and hospitalizations (Iezzoni & Ngo, 2007). In addition, O'Connell et al. (2014) found the largest component of indirect costs for an MS relapse was loss of income due to employment status. Relapses function independently of other clinical MS symptoms; however, relapses can worsen MS symptoms and accelerate disease progression (Koch-Henriksen, Thygesen, Sørensen, & Magyari, 2019).

Improvement science. MS has only recently entered the realm of improvement science and population health. Groundwork for this has been established by large epidemiological registries, such as the North American Research Committee on Multiple Sclerosis (NARCOMS), the Multiple Sclerosis Partners Advancing Technology and Health Solutions (MS PATH), and Multiple Sclerosis Leadership and Innovation Network (MS LINK). NARCOMS is a data registry that utilizes patient experience to improve clinical care and quality of life (NARCOMS, 2017). MS PATH uses technology and collects MS patient data from routine office visits to enable MS research (MS PATH, 2016). MS LINK is an interdisciplinary research community that is focused on improving the care of an MS individual (Gisler, 2019). Finally, the Slifka Longitudinal MS Study researches population demographic and clinical measures of utilization, illness, and cost (Minden et al., 2006).

Historically, MS care has been studied and improved at the basic and clinical trials research and epidemiological studies on outcomes and quality of life. These approaches have not studied MS care or outcomes at a system (MS Center/MS clinic) level. MS-CQI represents the vanguard of this approach utilizing an "improvement science approach." The improvement science approach combines QI with rigorous scientific methodology.

Continuous quality improvement. The QI approach is a system level approach influenced by the IHI, the IOM and the Affordable Care Act. The IHI triple aim specifically suggests a new systems-oriented focus and continuous improvement culture (IHI, n.d.). Also, the IOM shows significant quality and safety deficiencies in our current health system (Nationalacademies.org, 2019).

To Err is Human: Building a Safer Health System (IOM, 2001) was published in 1999. The authors stated that tens of thousands of people in the United States die every year due to

mistakes that are preventable with their care (IOM, 2001). This publication led to a call for a more systems-level change and *Crossing the Quality Chasm* (IOM, 2001) was published with a focus on how the U.S. health system needs to be overhauled to improve the delivery of care to PwMS (Fanjiang, Grossman, Compton, & Reid, 2005). The IOM suggests six aims of improvement: safe, effective, patient-centered, timely, efficient and equitable (IOM, 2001). Six redesign imperatives were notated to drive these areas of improvement: (a) reengineered care processes; (b) effective use of information technologies; (c) knowledge and skills management; (d) development of effective teams; (e) coordination of care across patient conditions; (f) use of performance and outcomes measurement (IOM, 2001). The LHS model was popularized by this work, proposing that improvement and research could be simultaneously conducted and better outcomes achieved by focusing on system-level performance (Nelson et al., 1995). MS-CQI was created as the first LHS improvement research collaborative for MS, and is addressing some of these deficiencies by utilizing a learning health system model that employs integrated improvement and research methods.

There are many popular QI approaches that are used nationally among CQI collaboratives such as the IHI Model for Improvement, Clinical Microsystems, and Lean Six Sigma (IHI, n.d.; Clinical Microsystems, n.d.; Lean Six Sigma Online, n.d.). These different methodologies are all utilized in healthcare settings across the United States and when implemented, can have significant results. The Clinical Microsystems approach focuses on “microsystems,” which are defined as the essential building blocks of larger health system (Clinical Microsystems, n.d.). Clinical microsystems complete value-added and hands-on work inside health organizations and are living units that have the patient and their needs, front and center (Clinical Microsystems, n.d.). This approach builds upon foundational work by business scholar P. Brian Quinn which

documented the competitive advantage of businesses that excelled in front-line service performance focused on optimizing the capability of "smallest replicable units" or SRUs—improvement of SRU performance translated into improved overall performance (Quinn, 1992). The SRU theory was later adapted for healthcare by Batalden and Nelson to develop the microsystems approach. This approach guided the systems-level focus for the MS-CQI study. (Nelson, Batalden, Godfrey, & Lazar, 2011).

Lean Six Sigma's primary goal is to reduce waste and variation to create an environment of optimal quality control (Lean Six Sigma Online, n.d.). This approach includes the use of SPC methods to study and improve performance variation – these include the use of SPC for benchmarking in improvement initiatives and collaboratives (Benneyan, Lloyd, & Plsek, 2003). This approach influenced the benchmarking approach used in the MS-CQI study. The "Model for Improvement" is used by healthcare organizations as a framework to accelerate healthcare improvement through iterative small tests of change aimed at optimizing but not replacing the existing model ("Institute for Healthcare Improvement: How to Improve," n.d.). The fundamental structure of the IHI Model for Improvement includes piloting the changes on a small-scale utilizing Plan-Do-Study-Act (PDSA) cycles, see Figure 1 ("Institute for Healthcare Improvement: How to Improve," n.d.). The IHI model is influenced by Kolb's experiential learning model for adult learning and in turn, has influenced modern design approaches including human centered design and agile prototyping development (Kolb, 1984). This approach influenced the design of the improvement intervention for the MS-CQI study, which includes professional improvement coaching and a modified version of the IHI Breakthrough Series (BTS) collaborative structure (IHI, 2003) to guide microsystem teams applying PDSA cycles to improve system performance.

System-level change focus. The current state of the healthcare system in the United States has recently moved towards a value-based quality of care (Porter, 2009). In particular, the Affordable Care Act is driving a shift from productivity to systems-level value-based reimbursement (healthcare.gov, 2019) and suggests that a new paradigm of continuous quality improvement will be required to optimize value and quality. It will be necessary to demonstrate the value of MS care at systems and population levels, especially in complex, high cost, chronic "3C" conditions such as MS.

De Silva (2014) reported that quality improvement collaboratives facilitate systems-level change in real-world settings. Regional and national CQI collaboratives that are utilizing these improvement methodologies have demonstrated significant results. The Northern New England Cardiovascular Network (NNE) Disease Study Group has used a shared data registry and QI methods to reduce mortality and morbidity across cardiac surgery centers in the U.S. ("Northern New England Cardiovascular Network Disease Study Group," n.d.). Research on quality collaboratives for hypertension has also been successful. Participating Medicaid managed care plans in California reduced the hypertension of their patients (Backman et al., 2017). There was an increase of 5% in their rates of controlled hypertension (Backman et al., 2017).

The Vermont Oxford Neonatal Health Network (VONHN) is a nonprofit collaboration of health care professionals as an interdisciplinary community to change the landscape neonatal care (Vermont Oxford Network, n.d.). The goal of VONHN is to improve the safety, quality, and value of care for newborn babies and their families through a collaboration program of education, quality improvement, and research (Vermont Oxford Network, n.d.). The Inflammatory Bowel Disease (IBD) Qorus is a collaboration between patients and healthcare

providers to improve quality care and improved outcomes for people with IBD (Crohn's & Colitis Foundation, n.d.).

The Cystic Fibrosis Foundation (CFF) Learning and Leadership Collaboratives (LLCs), where 110 centers have participated in a national QI collaborative for over a decade. The CFF utilizes QI methods as well as a systems-level registry. The CFF collaboratives have had success in reduced mortality, improved life expectancy, decreased morbidity, and developed several process quality indicators (Godfrey& Oliver, 2014; Marshall & Nelson, 2014; Mogayze et al., 2014; Sabadosa & Batalden, 2014).

The differentiating characteristic of MS-CQI is that it includes a randomized research design. The prior collaboratives are cohort designs such as VONHN or do not have research designs (CFF, IBD). The MS-CQI is the vanguard effort for improvement science in MS because it builds upon these prior exemplars by merging QI and randomized research approaches. Additionally, current evidence suggests that improvement collaboratives and improvement coaching are equally effective in achieving success in clinical outcomes (Gustafson et al., 2013). Also, Gustafson and colleagues (2013), discuss that combining both coaching and improvement collaboratives optimize results since the components are additive. The MS-CQI intervention is based on the utilization of an improvement coach in combination with an improvement collaborative model.

Parent study design. The parent study (MS-CQI) is utilizing a prospective step-wedge randomized design, which is also referred to as a Comprehensive Dynamic Trial (West et al., 2008). The step-wedge cluster randomized trial design has become increasingly popular within clinical trials, and can be used in the evaluation of service delivery type interventions (Hemming, Haines, Chilton, Girling, & Lilford, 2015). In addition, the step-wedge design allows for

comparisons within microsystems and can better accommodate smaller sample sizes (West et al., 2008). In a standard step-wedge design, clusters randomly and sequentially crossover from control to intervention until most or all of the clusters are exposed to the intervention. At the beginning of the trial, no clusters are exposed (Hemming et al., 2015). After a baseline period, cluster(s) are successively randomized in a step-wise fashion to go from control to intervention until most or all the clusters have been exposed (Hemming et al., 2015). Therefore, each cluster is part of the control and intervention, which allows robust scientific evaluation in shorter time periods and with fewer clusters than standard cluster-randomized designs (Hemming et al., 2015).

The intervention employed by MS-CQI is a modified IHI Breakthrough Series clinical quality improvement (CQI) intervention, which augments the traditional BTS model by adding professional improvement coaching to accelerate front-line improvement capability (IHI, 2003; Oliver, Messier, & Hall, 2019). The IHI Breakthrough Series, CQI intervention with Professional Improvement Coaching was coordinated by clinical teams at participating sites under the guidance of a professional improvement coach utilizing the IHI improvement collaborative model. This intervention is a hybrid adaptation of the LLC and CFF (Godfrey & Oliver, 2014). The general structure and process, as outlined by and the IHI Breakthrough Series improvement collaborative model, was utilized for this intervention (Kilo, 1998; see Appendix A).

The core MS-CQI study used descriptive statistics as well as inferential outcomes analysis to assess performance across the four centers and between exposure (QI) and control (usual practice) conditions. The SPC benchmarking with variation analysis by center is also being conducted. The proposed study built upon the parent study and focused on predictive

analytics and more advanced statistical techniques that can complement and improve upon current MS-CQI analysis.

Predictive statistical modeling. In MS, relapses have been a focal point of research; however, there is a dearth of research on predicting MS relapses. Research has been used to determine predictors of MS relapse rate utilizing Poisson regression using demographic, clinical, and MRI variables (Held, Heigenhauser, Shang, Kappos, & Polman, 2005). Held et al.'s (2005) study included 821 participants from the placebo group of the Sylvia Lawry Centre for Multiple Sclerosis Research (SLCMSR) and was aimed at determining prognostic factors available at baseline to the on-study relapse rate in MS. The timeframe of the study included 24 months prior to the study baseline and the MS disease course. The univariate analysis included 10 predictors for predicting the on-study relapse rate (Held et al., 2005). Results of the univariate analysis show that the on-study relapse rate was higher for younger female participants, for relapse-remitting participants versus secondary progressive participants, and for participants with positive enhancement status at entry using Wilcoxon test, $p < .05$ (Held et al., 2005). A higher relapse rate was also associated with a shorter disease duration, lower entry EDSS, more pre-study relapses, and more enhancing lesions in T1 (Held et al., 2005). A Poisson regression model was also fitted using five predictors to predict the on-study relapse rate (Held et al., 2005). The fitted Poisson model results showed that disease duration (estimate = -0.02) and previous relapse number were significant ($p < .05$) predictors (estimate = 0.59 for one, 0.91 for two, and 1.45 for three or more relapses vs no relapses; Held et al., 2005). While the Poisson regression is a solid traditional model for a continuous outcome, other sophisticated predictive modeling approaches could have been evaluated. In addition, only including 10 features in the model can be limiting in increasing accuracy and optimization.

Research has also been done on predicting RRMS utilizing discrete distribution models to characterize relapsing-remitting contrast-enhancing lesions (CEL) to quantify the interpatient variability (Velez de Mendizabal et al., 2013). Mendizabal et al.'s (2013) study analyzed nine MS participants over 48 months that had a monthly MRI. Sixteen structural models that included seven different probability distributions were evaluated with a maximum of six predictors (Velez de Mendizabal et al., 2013). Results showed that based on the number of model parameters and the precision of the parameters the best fitting model was the negative binomial compared to the other models such as Poisson model, Poisson model with mixed distribution, Zero-Inflated and Generalized Poisson models and the Zero-Inflated Negative Binomial model (Velez de Mendizabal et al., 2013). The minimum value of the objective function, which is approximate to -2 × log (likelihood) [-2LL], was criteria used as model comparison during model development (Velez de Mendizabal et al., 2013). The Akaike information criteria (AIC) was also used as measurement in the final model. In addition, the final model was based on the precision of parameter estimates and the results from model predictive performance from model simulations were compared (Velez de Mendizabal et al., 2013). While Velez de Mendizabal and colleagues' model compared 16 different structural models, the sample size was extremely small (n = 9) and results may not be generalizable to the MS population. Measurement on evaluating the best performing model used was a strength of the study, however, traditional statistical models were used and machine algorithms were excluded, which leaves a gap in the model evaluation of the study.

There is also some evidence on implementing a score to identify individual PwMS with RRMS via a multivariate Cox analysis, where the first relapse was the independent variable to try and predict future relapses in the short term (Sormani, Rovaris, Comi, & Filippi, 2007).

Sormani et al.'s (2007) study included 539 PwMS from the placebo group of a clinical trial. The univariate Cox analysis that produced variables with a *p* value less than .20 were used in the multivariate Cox analysis and time to first relapse was the dependent variable (Sormani et al., 2007). Variables that had a *p* value of .01 were retained in the final model. The final model produced a linear predictive score that was calculated using the variables included in the model and their estimated coefficients (Sormani et al., 2007). Creating a linear predictive score using a multivariate Cox analysis is a traditional statistical method. The study lacks comparison to other traditional and machine learning algorithms to evaluate fit and performance. In addition, the final model has only two independent predictors of relapse, the number of enhancing lesions on a baseline MRI and the number of relapses during the previous two years (Sormani et al., 2007). The number of features included in the final model may need to be expanded to generate a more accurate predictive model.

There is a lack of research on predicting MS hospital utilization, specifically MS all-cause UC. A recent study forecasted MS hospitalization in the U.S. from 2017-2040, and utilized an Autoregressive Integrated Moving Average (ARIMA; Sharma, Bittner, & Cho, 2019). Results showed hospital admissions were predicted to increase by 32% in 2030 (Sharma et al., 2019). However, this study does not compare with other sophisticated predictive models, and does not predict all-cause UC utilization. In 2005, a cross-sectional study of 4,000 MS participants was done to estimate current costs and quality of life of participants treated with disease-modifying drugs (DMDs) in the U.S, and an overall assessment of MS cost (Kobelt, Berg, Atherly, & Hadjimichael, 2006). Results showed that 28.8% of participants had a relapse during the past three months and total average costs were estimated at $47,215 per patient per year (Kobelt et al., 2006). Iezzoni and Ngo (2007) completed a study and interviewed 983 MS working-age

individuals. Results of the study showed significant impact on cost of MS care and disability impact. Of those interviewed, 27.4% indicated that, their health insurance concerns had significantly affected employment decisions since being diagnosed with MS (Iezzoni & Ngo, 2007). Furthermore, 27.4% put off or postponed seeking needed health care because of costs, 16.4% reported difficulty paying for health care, and 22.3% delayed filling prescriptions, skipped medication doses, or split pills because of costs (Iezzoni & Ngo, 2007). In addition, 26.6% reported significant concerns about affording food, utilities, and housing (Iezzoni & Ngo, 2007).

Overall, there has been limited research on predicting MS relapses and predicting MS all-cause UC utilization to date. These studies used traditional statistical methods with no emphasis on fit or performance of the model. In addition, all of the statistical methods used were traditional in nature and did not use modern predictive analytics techniques such as random forest, ridge or LASSO regression which will likely perform better and provide more robust and accurate results.

Statistical modeling comparisons. The main purpose of this research study is to compare statistical model classes between traditional methods and more sophisticated predictive analytics. There is some research suggesting what are the “best” model comparison approaches and how do they compare. Couronné, Probst, and Boulesteix (2018) conducted a model comparison paper between random forest and logistic regression. This was a large-scale benchmarking experiment based on 243 datasets comparing the prediction performance of random forest with logistic regression (Couronné et al., 2018). This model comparison used AIC, area under the curve (AUC), and Brier score and showed the delta between logistic regression and random forest. Results of the study showed that random forest performed more favorably over logistic regression in 67 of the datasets analyzed (Couronné et al., 2018). Mean differences

between logistic regression and random forest was .03 for accuracy, -.03 for Brier score, and .04 for AUC. Couronné et al.'s (2018) comparison study was utilized as a framework for this proposed study.

Zhang, Lu, and Hou (2019) conducted research on a model comparison between random forest and logistic regression in predicting diabetes. The study calculated AUC, sensitivity and accuracy for model comparison. Results showed that the accuracy of random forest was slightly higher than logistic regression (0.87 versus 0.86; Zhang et al., 2019). The AUC was similar between random forest and logistic regression and random forest (0.94 vs. 0.93; Zhang et al., 2019). This study is limited as the researchers only compared a few different classification measures and excluded measures such as precision, F1 Score, and Matthew's correlation coefficient (MCC) which this study has built upon (Zhang et al., 2019).

Matthew's correlation coefficient was originally developed in 1975 for comparison of chemical structures (Matthews, 1975). The MCC is similar to computing any correlation coefficient, but for two binary variables, true class and predicted class (Baldi, Brunak, Chauvin, Andersen, & Nielsen, 2000). The MCC is always between -1 and +1, a value of -1 denotes total disagreement and +1 total agreement (Baldi et al., 2000). Baldi et al. (2000), re-proposed Matthew's correlation coefficient as the standard performance metric for machine learning. Chicco and Jurman's (2020) study utilized six synthetic use cases in different genomics scenario, and showed MCC produced a more informative and accurate score in assessing binary classification models, than accuracy and F1 score. The F1 score is defined as the harmonic mean of precision and recall (Chicco & Jurman, 2020). The study suggests that the MCC should be preferred to F1 and accuracy when evaluating binary classification tasks by scientific communities (Chicco & Jurman, 2020). When a dataset is unbalanced, accuracy is not a good

measurement because it produces an over optimistic estimation of the classifier ability on the majority class, but MCC performs better in this situation (Chicco & Jurman, 2020). Thus, the MCC is a reliable statistical measure and produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (Chicco & Jurman, 2020). The confusion matrix is a critical tool used to determine the model's accuracy, is part of the model's results, and contains four metrics (true positives, false negatives, true negatives, and false positives).

There is a tradeoff between prediction accuracy and model interpretability when comparing parametric and nonparametric models in statistical learning methods. More restrictive models are generally more interpretable but sometimes less accurate in prediction (e.g., linear or logistic regression; James et al., 2013). If inference is the goal, then more restrictive models such as linear regression are usually preferred. Non-parametric methods such as decision trees and support vector machines are usually less flexible models and are less interpretable, but sometimes more accurate (James et al., 2013). With regards to this study, logistic, LASSO, and ridge are parametric methods and random forest is nonparametric. Figure 2 shows the flexibility of a method increases as the interpretability decreases (James et al., 2013). In addition, machine learning algorithms have the ability to include many features and perform optimally, where logistic regression may take more time to engineer and optimize the model.

Current research on relapse and all-cause UC does not enable broad and modern use of predictive modeling techniques. In addition, there is no research comparing fit and performance between predictive analytics and statistical methods. These limitations present challenges for cost-effectiveness analyses and the overall assessment of the quality and cost of MS healthcare services delivery. These studies have made the advancement of analysis of MS at the population

level. However, they have not conducted systems-level analysis nor investigated predictive analytics techniques concerning MS relapse or all-cause UC.

There are healthcare quality outcomes and cost implications for this work beyond the methodological comparisons of statistical models in theoretical research. Predictive analytics methods have the potential to empower and accelerate the capability of LHS to predict and respond to emerging health needs of people with costly, complex, chronic "3C" conditions like MS and better inform continuous improvement efforts. This could potentially result in achieving more efficient and effective healthcare outcomes and saving time and overall cost.

Chapter 2

Method

Parent Study Design

The parent research study, MS-CQI (Appendix A), utilizes a prospective step-wedge randomized design. Compared to a standard randomized clinical trial or cluster-randomized design, the step-wedge design can allow for additional comparisons within each MS center in a pre-clinical and post-clinical quality intervention (Rapkin et al., 2012). Randomized clinical trials are the gold standard in research design for inferring a causal relationship (Cartwright, 2010).

MS-CQI is a two-part, 3-year clinical quality improvement prospective study that started in June 2017 and concluded in June 2020. There are two levels of participation in this study: (a) system-level administrative; and (b) individual-level clinical Patient Reported Outcomes (PROs). In MS-CQI, the term "system level" refers to the MS clinical sites and "population level" refers to the whole MS collaborative which consists of all four clinical sites aggregated together. Individual level data are available within each MS clinical center. These data created a combined MS clinical quality intervention (MS-CQI) database.

Benchmarking and analysis of data from each center has been conducted via the MS-CQI database quarterly throughout the three-year study. Baseline data were collected during the first year of the study. In the second part of the study, years two and three, investigations were conducted on the effect of a clinical quality intervention on primary endpoints and selected secondary measures at the system and individual level outcomes. Protected health information (PHI) was collected for the individual level PRO portion of the study; therefore, written informed consent was obtained for participation in this part of the study. (Oliver et al., 2019).

Proposed Study Design

This study was built on the standard MS-CQI protocol which has established the MSCQI systems-level database and utilized data from the parent study for secondary analyses. An analytic comparison study was proposed based on the third year of data, June 2019-June 2020. The MS-CQI team has studied system-level variation in relapses and all-cause UC for individual sites, between sites, and for MS-CQI collectively as part of the parent study. The proposed study predicted relapse and all-cause UC and compares these outcomes through various types of statistical models. Data collected across four clinical MS centers include administrative data and eight clinical electronic health record (EHR) clinical outcome measures. These outcome measures are described in the subsequent section. The MS-CQI study is using traditional methods of analysis, and this study builds upon that by studying newer approaches and comparing these with the standard approaches. Logistic regression and maximum likelihood estimation methods were used for standard inferential analyses. Ridge, LASSO, and random forest was used for predictive analytics.

Participants

This study used EHR data extracted at the individual level and 'rolled up' to the system and population level. Inclusion criteria included participants age 18 years or older, with MS presenting to any of the four centers, who entered the study in any quarter (7/1/2017-6/30/2020). Exclusion criteria included cases with missing or incorrectly input data and those who refused to participate in the study. For the purpose of this study, data collected in year 3 was utilized. Additional methods for participants are available in the parent study (See Appendix A).

Setting

Methods for setting are described by the parent study protocol for MS-CQI (see Appendix A).

Data Management

Data were extracted in a de-identified form via data downloads from electronic medical records (EMRs) and administrative records from participating centers. Each MS-CQI clinical center enters individual level data, and the data was managed using REDcap (Research Electronic Data Capture) electronic data capture tools hosted at Dartmouth. REDCap is a secure, web-based software platform designed to support data capture for research studies, providing (a) an intuitive interface for validated data capture; (b) audit trails for tracking data manipulation and export procedures; (c) automated export procedures for seamless data downloads to common statistical packages; and (d) procedures for data integration and interoperability with external sources (Harris et al., 2019; Harris et al., 2009). No PHI was entered into this database. All PHI remained at specific clinical center locations. All data were initially managed in the REDcap database, and subsequently extracted into SAS 9.4 for data management, and R for statistical analysis. Additional methods for data management were established by the MS-CQI parent study (See Appendix A).

Procedures

Recruitment. Methods for recruitment were established by the MS-CQI parent study (See Appendix A).

Informed consent. The institutional review board (IRB) at Dartmouth (Committee for Protection of Human Subjects at Dartmouth College) approved the parent MS-CQI study, and a reliance agreement between Dartmouth and the Human Research Protection Program at the

University of Indianapolis was put into place before data analysis was conducted. Additional methods for informed consent were established by the MS-CQI parent study (See Appendix A).

Continuous Quality Improvement Intervention

The clinical quality improvement intervention employed in the parent study is an IHI Breakthrough Series CQI intervention, including an improvement collaborative and professional improvement coaching (Oliver et al., 2019). Methods for CQI intervention were established by the MS-CQI parent study (See Appendix A).

Data and Data Collection for Relapses and All-Cause Hospitalization

Each of the four MS-CQI clinical centers abstracted participant data from their own EHRs. The data were deidentified and then input into REDcap. The following is an estimated sample size for unique participants by center for EHR year 3: Center A: $n = 977$; Center B: $n = 460$; Center C: $n = 539$; Center D: $n = 556$. Total sample size for year 3 is 2,532. For this study, data was extracted from REDcap. Data management for this study used SAS/STAT® software, Version 9.4 (Copyright ©2019; SAS Institute Inc., Cary, NC). Analysis for this study was conducted in R (R Core Team, 2014). The following variables were extracted and included in the predictive models.

- Patient level predictors include:
    - Age is a continuous variable measured in years
    - Sex is a categorical dichotomous variable (male = 0, female =1)
    - MS Center is an ordered categorical variable (1, 2, 3, 4)
    - MS phenotype is an ordered categorical variable (RRMS, SPMS, PPMS, PRMS, other).
- Clinical level predictors include:

- Participants on at least one DMT is a dichotomous variable (0 = no, 1 = yes)
- Participants with at least one brain MRI is a dichotomous variable (0 = no, 1 = yes)
- Participants with at least one thoracic MRI is a dichotomous variable (0 = no, 1 = yes)
- Participants with at least one cervical MRI is a dichotomous variable (0 = no, 1 = yes)
- Participants with at least one MS relapse is a dichotomous variable (0 = no, 1 = yes)
- Participants with at least one hospitalization is a dichotomous variable (0 = no, 1 = yes)
- Participants with at least one ED visit is a dichotomous variable (0 = no, 1 = yes)
- Participants with at least one UC visit is a dichotomous variable (0 = no, 1 = yes)

Exploratory Data Analysis

Descriptive statistics include frequency distributions for categorical variables and means with standard deviations for normally distributed continuous variables. The output of the descriptive statistics was reported in tabular form for each center and overall. Appropriate power tests were conducted to ensure at least a .80 level, and demographic data were analyzed for differences across centers. Mean differences were tested via one-way ANOVA (analysis of variance) for continuous variables and chi-square for categorical variables with a significance level of $\alpha = .05$. Bivariate analysis was conducted to determine associations between pairs of variables and which variables should be included in the predictive models ($p < .10$). If there were missing data, an average imputation was computed, and the data were included in the analysis.

Boxplots were utilized to assess outliers in continuous data and frequency tables for categorical data. If extreme observations were detected, the analysis continued to better understand if the outliers impacted the analysis results. If the results were affected by the outliers, those observations were removed from the dataset.

Predictive Models

Logistic regression was utilized as the traditional model because it is often used for classification predictive models in healthcare. Its popularity in health sciences is due to the ability of logistic regression to give a discrete dichotomous outcome (e.g., disease or no disease; Tabachnick, Fidell, & Ullman, 2019). In the proposed study, logistic regression was compared with other machine learning models, specifically ridge, LASSO, and random forest. The area under the curve of these models was compared to assess which model had the best goodness of fit to suggest which approach is best at predicting MS relapse and MS all cause-UC utilization. In addition, accuracy, true positive rate (TPR), false-positive rate (FPR), negative predictive value (NPV), precision (PPV), F1 measure, and MCC was compared across models. Each model was separately estimated for each outcome, and model results were tabled and compared.

Logistic regression. Logistic regression modeling allows a discrete outcome variable (dependent variable) to be modeled as a function of other variables that can be discrete, continuous, dichotomous, or a mix (Tabachnick et al., 2019). Logistic regression analysis is also used when the distribution of responses on the dependent variable is likely to be nonlinear (Tabachnick et al., 2019). The logistic regression is a nonlinear model and the equations utilized to describe the outcomes are more complex than those for multiple regression (Tabachnick et al., 2019). However, this nonlinear function is based on the best linear combination of predictors (Tabachnick et al., 2019). This linear equation uses the logit function as our link function and is

the natural ($\log_e$) of the probability of being in one group divided by the probability of being in the other group (Tabachnick et al., 2019). The technique for estimating coefficients in logistic regression is maximum likelihood. The purpose of this procedure is to find the best linear combination of predictors to maximize the likelihood of obtaining the observed outcome frequencies (Tabachnick et al., 2019). Since the logit function is utilized, the output is the log-odds; thus, we call this model logistic regression. The range of log-odds is from $-\infty$ to ∞. In this study the logistic model was optimized initially through the bivariate analysis in exploratory data analysis (EDA). In the modeling process the variables that were selected through the EDA were run through backward elimination at the statistical significance of $p < .05$. This process engineered and optimized the logistic regression to compete equitably with the machine learning algorithms.

Ridge regression and LASSO. Ridge regression and LASSO are both shrinkage methods and can be used with many features in the model without overfitting (Hastie, Tibshirani, & Friedman, 2009; James et al., 2013). Both of these models shrink or regularize the coefficient estimates ("shrinks" them) towards zero (Hastie et al., 2009; James et al., 2013). By shrinking coefficient estimates, this improves the fit of the model by reducing the model's overall variance (Hastie et al., 2009; James et al., 2013). Ridge and LASSO work by adding a penalty term to the log-likelihood function.

For ridge regression, the penalty term is β_j^2, and for LASSO, it is $|\beta_j|$. (Hastie et al., 2009; James et al., 2013). The penalty term is a shrinkage penalty, and when the beta estimates are small it has the effect of shrinking the parameters close to zero (James et al., 2013). The tuning parameter controls the relative impact of the shrinkage penalty on the coefficient estimates. When the tuning parameter, λ, is positive, the coefficient estimates of large values of β are

shrunk towards zero. When $\lambda = 0$, the penalty term has no effect, and the ridge regression will produce least squares estimates (Hastie et al., 2009; James et al., 2013). The value of λ is critical and can be chosen via cross-validation (James et al., 2013). Cross-validation to be discussed subsequently.

Ridge will never force any variables to zero, so every variable is retained in the final model, which affects the interpretability of the model (Hastie et al., 2009; James et al., 2013). The penalty term makes the ridge regression model have decreased variance but increased bias (Hastie et al., 2009; James et al., 2013). In this circumstance, bias refers to how far off our predicted value is compared to the actual value. Variation is the spread or variation in the predicted values. The key in machine learning is to find a balance between bias and variation. For example, in machine learning the model is often complex, as complexity increases, variance increases and bias decreases. Finding the perfect balance between bias and variance increases the accuracy and generalizability of the model. Figure 3 shows the ridge regression coefficient estimates for each value of λ. As λ increases, the coefficient estimates approach zero (the horizontal line) and reduces the complexity of the model (James et al., 2013).

LASSO works similarly to ridge but has an alternative penalty term of $|\beta_j|$. The LASSO penalty term will set some variables to zero that are not important to the outcome; this occurs when the tuning parameter, λ, is sufficiently large and therefore performs variable selection (James et al., 2013). When $\lambda = 0$, the resulting model is an ordinary least squares (OLS) model. When λ is very large, the resulting model is the null model (intercept-only model) as all variables are forced to 0 (James et al., 2013). LASSO will produce a model that is easy to interpret due to variable selection and with high predictive power (Hastie et al., 2009; James et al., 2013). Ridge

regression and LASSO are similar in that as the tuning parameter λ increases, so does bias, but variance decreases (Hastie et al., 2009; James et al., 2013).

Random forest. Decision trees are used frequently because they are similar to the way human beings make decisions and thus are easier to explain and visualize (James et al., 2013). However, their predictive accuracy may not be as strong as other machine learning predictive models (James et al., 2013). Random forest is a decision tree-based algorithm that does not have overfitting issues as most decision-tree algorithms do, and increased predictive accuracy (Hastie et al., 2009; James et al., 2013). This is because random forest is different from other decision trees; it decorrelates the trees (Hastie et al., 2009; James et al., 2013). The decorrelation process is also called bagging, which generates new training data sets from an original data via random sampling with replacement, also known as bootstrapping. This tree model uses various sets of training data, but each time there is a split in the tree a random sample of predictors are chosen from the full set of predictors (Hastie et al., 2009; James et al., 2013). Each individual bootstrapped data set is then used to construct a tree within the forest. This methodology reduces overfitting the model and is far less likely than other decision tree models to have a small set of strong predictors in all the tree splits (Hastie et al., 2009; James et al., 2013).

Cross-validation. Resampling methods are an invaluable tool in the era of modern statistics (James et al., 2013). Resampling involves drawing samples repeatedly from the training set and refitting the model on each of the given samples to learn additional information on the fitted model (James et al., 2013). This approach allows us to learn new information about the model by fitting it multiple times and comparing different sample results as opposed to just fitting it once in the training sample (James et al., 2013). Data for the MS-CQI study are highly unbalanced-- particularly the relapse outcome. The low prevalence causes an imbalanced dataset

as there are far many more non-events (no relapse) than events (relapse). During cross-validation, a synthetic minority over-sampling technique (SMOTE) algorithm was used to help with the imbalanced nature of the data. This algorithm is well known and is used to synthetically generate new examples of the minority class via nearest neighbors and the majority class, in this case (no-relapses) are also under-sampled, for a more balanced dataset. (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). Ridge, LASSO, and random forest methods generate a variable importance plot, which is scored from 0 to 100 and gives a value based on the statistical significance and effect on the model

Cross-validation is a type of resampling method that is used to estimate the test error of the statistical method, which gives information on model performance or flexibility (James et al., 2013). Cross-validation can be a useful approach in classification models where the outcome Y is qualitative (James et al., 2013). In this situation, the cross-validation uses the number of misclassified observations (James et al., 2013). For this study *k*-fold cross-validation was used as the resampling method, where $k = 10$. The k-fold approach randomly divides the observations in the data set into folds or *k* groups of equal size (Hastie et al., 2009; James et al., 2013). The initial fold is the validation set and then the method is fit on the remaining k-1 folds (Hastie et al., 2009; James et al., 2013). This process is then repeated *k* times with a different set of data and each time is treated as a validation set (James et al., 2013). The standard approach for classification problems in data science is 10-fold cross-validation.

For this study, the best model from each cross-validation training of 10 models is selected as the champion (optimized) model. During cross-validation each model was separately optimized for internal tuning using accuracy, MCC, and F1. For the endpoint of relapse, logistic regression, LASSO, ridge and random forest were each cross-validated ten times by three

different optimizations for a total of 12 champion models that came from a total of 120 models. For the endpoint of all-cause UC logistic regression, LASSO, ridge and random forest each was cross-validated ten times by two different optimizations (accuracy and MCC) for a total of eight champion models that came from a total of 80 models. The additional optimization method of F1 was not needed for all-cause UC as the dataset was already sufficiently balanced and produced similar results for accuracy and MCC.

Model comparison. The AUC, accuracy, TPR, FPR, precision, NPV, F1 score, and MCC was computed for all models. In statistical tests with binary outcomes, the accuracy is evaluated by sensitivity and specificity. Sensitivity also known as "recall" is defined as true positives, and specificity is defined as true negatives. Plotting the sensitivity versus 1- specificity generates the receiver operating characteristic (ROC) curve (Hajian-Tilaki, 2013). The AUC is an effective measure of this accuracy in predictive models (Hajian-Tilaki, 2013). Generally, the AUC range is between .50 and 1.0, and is also known as the "c-statistic." An AUC of .50 indicates a prediction by chance, and AUC of 1.0 indicates a perfect prediction (Tabachnick et al., 2019). Accuracy is an overall average of how well the model predicts as well as its computational simplicity. In data science positive predictive value (PPV) and precision have the same mathematical definition, however the actual nomenclature is different. Precision or PPV is defined as true positives divided by true positives plus false positives. The precision or PPV defines the probability of having the disease or state in an individual with positive result (Šimundić, 2009). The F1 score is defined as the harmonic mean of precision and recall (Chicco & Jurman, 2020). The MMC is not affected by the unbalanced datasets, and is a contingency matrix method of calculating the Pearson product- moment correlation coefficient between actual and predicted values (Chicco, & Jurman, 2020).

Model metrics in Table 1 and Table 2 were calculated for all models in this study and ridge, LASSO, and random forest were compared to logistic regression. Building upon the framework established by Couronné and colleagues (2018), the delta between AUC, MCC, and F1 Score were added and compared for each model pair. This combination of measures was chosen based on model comparison studies in the literature, as well as MCC being the "new" measure to evaluate machine learning model's performance and fit (Chicco & Jurman, 2020; Couronné et al., 2018; Zhang et al., 2019). Historically, F1 score and AUC were standards in comparing model performance and fit.

Indices were created and differences were calculated as follows*:*

- $\Delta perf_A = perf_{RF} - perf_{LR}$
- $\Delta perf_{A1} = perf_{LASSO} - perf_{LR}$
- $\Delta perf_{A2} = perf_{RR} - perf_{LR}$
- $\Delta perf_M = perf_{RF} - perf_{LR}$
- $\Delta perf_{M1} = perf_{LASSO} - perf_{LR}$
- $\Delta perf_{M2} = perf_{RR} - perf_{LR}$
- $\Delta perf_F = perf_{RF} - perf_{LR}$
- $\Delta perf_{F1} = perf_{LASSO} - perf_{LR}$
- $\Delta perf_{F2} = perf_{RR} - perf_{LR}$

Before beginning the analysis and to evaluate internal validity, a training and test data set were split into 70% and 30% test for validation. All models in this study followed the same validation method. Unless noted, the threshold for statistical significance is $p < .05$. All models have the same dataset, predictors, and outcomes, and all models were trained and fit on the entire data set of the backwards logistic regression model for consistency. Several data visualizations

were implemented as part of this study, including ROC and gain curves. Gain curves were used to visually evaluate model performance in a binary predictive model. The visualization presents the percentage of captured positive responses as a function of a selected percentage of the sample.

The following R packages were used in this study:

- tidyverse: package for data import
- caret: package for modeling and machine learning
- yardstick: package used to quantify model fit and performance
- glmnet: package for generalized linear models
- broom: package for data manipulation and analysis
- dplyr: package for data transformation and manipulation
- ggplot2: package for data visualization

Chapter 3

Results

Center and Participant Characteristics

Characteristics of participating MS-CQI centers are given in Table 1 for Year 3 of the study. The four participating centers (1-4), represent varying contexts ranging from urban academic MS centers to a private practice setting. Only two participants were dropped due to missing data. A total of 2,532 unique PwMS were followed by MS-CQI in the third year of the study, and volume varied substantively across centers from a low of 460 in center 2 to a high of 977 in center 1. Overall, the general demographic characteristics of the study population align with those established in general MS populations, including a majority of female gender (76%), and relapsing MS subtype (81%). Mean age was 50 years (see Figure 4). A number of important characteristics varied significantly ($p < .001$) across centers, including MS diagnosis type, age, and gender.

Clinical Outcomes and Center Level Variation

Chi-square tests identified numerous differences in center level performance outcomes (Table 2). Approximately 73% of participants were on disease modifying therapy (DMT), and center-specific performance ranged from 60% to 95% ($p < .001$). Overall brain (head) MRI utilization was 48% with variation observed in center level performance (37-56%, $p = .051$). Overall cervical (upper neck) and thoracic (middle back) MRI utilization was much lower (27% and 15%, respectively), with significant center level variation again observed in the cervical range: 18-35%, ($p < .001$); thoracic range: 4-23%, ($p < .001$). Approximately 4% of participants experienced an MS relapse and center-specific performance ranged from 2.6% to 6.5% ($p < .001$). The ANOVA procedure was used to identify significant differences on annualized rates

for ED and UC Utilization, relapse rate, and age across centers. All the measures varied significantly ($p < .001$; see Table 2).

The proportion of PwMS with at least one episode of acute care utilization (UC, ED, and hospitalizations) were relatively low: UC = 2.7%, ED = 8.2%, and hospitalization = 7.0%. Significant center level variation in performance was observed here as well. The proportions of PwMS with at least one hospitalization ranged from a high of 10.7% to a low of 4.3% ($p < .001$), proportions of PwMS with at least one ED visit ranged from a high of 11.1% to a low of 5.8% ($p < .001$), and proportions of PwMS with at least one UC visit ranged from a high of 5.2% to a low of 0.2% ($p < .001$). Utilization in these categories was also calculated in terms of utilization rates, and demonstrated statistically significant variation ($p < .001$) across centers (see Table 2).

A Spearman correlation showed .98 - .99 positive correlation between clinical level dichotomous variables of ED, UC, hospitalization and relapse and the associated continuous level utilization variables. Therefore, the continuous level utilization variables were removed from any further modeling and analysis.

Training and Testing Datasets

The main dataset was spilt into a 70% training dataset that included 1,773 participants and a test dataset that included 30% of the cohort or 759 participants. Detailed information on the training and testing cohorts are included in Table 3. Bivariate analysis for determining predictors included in the predictive models is available in Table 4.

Quantitative Aim 1: Relapse Predictive Model Comparison

The first aim of the study was to compare logistic regression to LASSO, ridge and random forest in predicting MS relapse. Relapse has very low prevalence. There was a total of 97 relapses in the dataset. The training data set had 74 relapses or 4.2% and the test dataset had 23

or 3.1% relapses. The low prevalence of relapse causes an imbalanced dataset, as there are far many more non-events than events. An initial backwards logistic regression was fit on the full dataset to set up the training and fit for consistency on all models. Based on the bivariate analysis and full dataset results the following features were included in all relapse models: age, gender, MS center, MS phenotype, DMT, ED visit, UC visit, brain MRI, cervical MRI, and thoracic MRI. Relapse was set as the target variable (outcome). There were 120 models run via cross-validation for three different optimizations (accuracy, MCC, F1) and four different models (logistic, LASSO, ridge, random forest) within each paradigm. Twelve champion models were selected for the endpoint (relapse).

Logistic regression (accuracy). A logistic regression using backward elimination was conducted controlling for confounding effects of age, gender and center. Center, MS phenotype, ED visit, brain MRI, thoracic MRI were significant at the $p < .05$ level. UC visit was not significant at $p = .07$. The confusion matrix predicted 582 true negatives, 18 false positives, 148 false negatives and 11 true positives. Some key measures of the model were accuracy = 0.78, AUC = 0.63, TPR = 0.38, FPR = 0.80, MCC = 0.08 and F Score was 0.12. Additional model performance measures are in Table 6.

LASSO (accuracy). A LASSO was conducted controlling for confounding effects of age, gender and center. The model produced an intercept only model due to shrinking the rest of the parameter estimates to zero. The confusion matrix predicted 730 true negatives, 29 false positives, 0 false negatives and 0 true positives. The model predicted no relapses on all observations, which is why certain performance measures such as MCC and F1 score could not be calculated. Some key measures of the model were accuracy = 0.96, AUC = 0.50, TPR = 0, and FPR = 1. Additional model performance measures are in Table 7.

Ridge (accuracy). A ridge regression was conducted controlling for confounding effects of age, gender and center. The model's variable importance was as follows: thoracic MRI (100.0), brain MRI (84.0), center 4 (82.9), UC visit (69.5), RRMS (60.5), SPMS (59.8), DMT (57.6), cervical MRI (54.5), ED visit (52.7), center 2 (50.5), center 3 (38.6), other phenotype (20.8), gender (4.5), Age (0.77) and PRMS (0). The confusion matrix predicted 730 true negatives, 29 false positives, 0 false negatives and 0 true positives. The model predicted no relapses on all observations, which is why certain performance measures such as MCC and F1 score could not be calculated. Some key measures of the model were: accuracy = 0.96, AUC = 0.67, TPR = 0, and FPR = 1. Additional model performance measures are in Table 8.

Random forest (accuracy). A random forest was conducted controlling for confounding effects of age, gender and center. The model's variable importance was as follows: thoracic MRI (100), center 4 (83.0), brain MRI (69.0), DMT (40.1), cervical MRI (39.5), ED visit, (35.2), gender (35.0), center 2 (34.7), center 3 (27.9), RRMS (26.9), Age (18.5 years), other phenotype (14.4), SPMS (13.3), UC visit (3.3) and PRMS (0). The confusion matrix predicted 677 true negatives, 20 false positives, 53 false negatives and 9 true positives. The model predicted no relapses on all observations. Some key measures of the model were: accuracy = 0.90, AUC = 0.69, TPR = 0.31, FPR = 0.93, MCC = 0.17 and F1 Score was 0.20. Additional model performance measures are in Table 9.

All relapse models (accuracy). Evaluating models for AUC, random forest performed the best at (0.69), ridge (0.67), logistic regression (0.63), and LASSO (0.51) (see Figure 5). Random forest AUC was 9.5% higher than logistic regression. LASSO and ridge models predicted no relapses for all observations therefore; MCC and F1 Score were not calculated. MCC for random forest was 0.17 versus 0.08 for logistic, a 100% difference. The F1 score was

0.198 for random forest versus 0.12 for logistic regression, a 69% difference. Figure 6 shows the gain curves across models and upon visual inspection random forest and ridge models need to test about 25% of the sample population to capture 50% of positive responses however logistic requires 35% and LASSO 50%. Additional model performance measures are in Table 22.

LASSO (MCC). A LASSO regression was conducted controlling for confounding effects of age, gender and center. The model's variable importance was as follows: thoracic MRI (100.0), center 4 (79.0), UC visit (61.5), RRMS (42.2), brain MRI (39.6), center 2 (36.6), gender (25.5), ED visit (22.1), center 3 (21.2), other phenotype (7.0), Age (0.72), cervical MRI (0), SPMS (0), DMT (0) and PRMS (0). The confusion matrix predicted 575 true negatives, 17 false positives, 155 false negatives and 12 true positives. Some key measures of the model were accuracy = 0.77, AUC = 0.62, TPR = 0.41, FPR = 0.79, MCC = 0.09 and F1 score was 0.12. Additional model performance measures are in Table 10.

Ridge (MCC). A ridge regression was conducted controlling for confounding effects of age, gender, and center. The model's variable importance was as follows: PRMS (100.0), thoracic MRI (98.2), UC visit (78.4), center 4 (75.4), brain MRI (67.3), center 3 (52.0), ED visit (49.9), SPMS (46.5), cervical MRI (45.4), center 2 (44.4), RRMS (43.50), DMT (37.2), gender (21.3), other phenotype (15.4), and age (0). The confusion matrix predicted 603 true negatives, 17 false positives, 127 false negatives and 12 true positives. Some key measures of the model were accuracy = 0.81, AUC = 0.65, TPR = 0.41, FPR = 0.83, MCC = 0.12 and F1 Score was 0.14. Additional model performance measures are in Table 11.

Random forest (MCC). A random forest was conducted controlling for confounding effects of age, gender and center. The model's variable importance was as follows: thoracic MRI (100.0), center 4 (98.0), brain MRI (67.6), ED visit, (43.8), DMT (41.1), center 2 (33.1), RRMS

(32.0), Age (31.5), gender (27.3), cervical MRI (26.2), center 3 (25.5), SPMS (20.5), other phenotype (10.7), UC visit (1.7) and PRMS (0). The confusion matrix predicted 669 true negatives, 19 false positives, 61 false negatives and 10 true positives. Some key measures of the model were accuracy = 0.90, AUC = 0.69, TPR = 0.35, FPR = 0.92, MCC = 0.17 and F1 Score was 0.20. Additional model performance measures are in Table 12.

All relapse models (MCC). For the MCC optimization, logistic regression results were the same as accuracy. Evaluating models for AUC, random forest performed the best at (0.69), ridge (0.65), logistic regression (0.63), and LASSO (0.62) (See Figure 7). Random forest AUC was 9.5% higher than logistic regression. MCC for random forest was (0.17), ridge (0.12), LASSO (0.09) and (0.08) for logistic regression. The F1 score for random forest was (0.20), ridge (0.14), LASSO (0.12) and (0.12) for logistic regression. Random forest outperformed logistic regression on MCC and F1 score by 107% and 71% respectively. LASSO and ridge also outperformed logistic regression. Figure 8 shows the gain curves across models and upon visual inspection all models need to test about 30-35% of the sample population to capture 50% of positive responses. Additional model performance measures are in Table 22.

LASSO (F1). A LASSO regression was conducted controlling for confounding effects of age, gender and center. The model's variable importance was as follows: thoracic MRI (100) the rest of the features had values of 0. The confusion matrix predicted 730 true negatives, 29 false positives, 0 false negatives and 0 true positives. Some key measures of the model were accuracy = 0.96, AUC = 0.59, TPR = 0, and FPR = 1. The model predicted no relapses on all observations, which is why certain performance measures such as MCC and F1 score could not be calculated. Additional model performance measures are in Table 13.

Ridge (F1). A ridge regression was conducted controlling for confounding effects of age,

gender and center. The model's variable importance was as follows: thoracic MRI (100), center 4 (92.6), brain MRI (68.4), UC visit (67.6), DMT (58.2), center 2 (54.5), cervical MRI (51.1), center 3 (47.8), RRMS (45.3), SPMS (40.5), ED visit (23.6), gender (13.2), other phenotype (4.6), Age (1.2) and PRMS (0.0). The confusion matrix predicted 654 true negatives, 17 false positives, 76 false negatives and 12 true positives. Some key measures of the model were accuracy = 0.88, AUC = 0.65, TPR = 0.41, FPR = 0.90, MCC = 0.19 and F1 score was 0.21. Additional model performance measures are in Table 14.

Random forest (F1). A random forest was conducted controlling for confounding effects of age, gender and center. The model's variable importance was as follows: center 4 (100.0), thoracic MRI (77.9), brain MRI (53.6), Age (42.4), ED visit, (36.6), cervical MRI (32.9), DMT (17.4), gender (15.3), center 2 (13.4), center 3 (7.4), SPMS (5.5), other phenotype (4.97), PRMS (1.19), UC visit (1.16) and RRMS (0). The confusion matrix predicted 645 true negatives, 23 false positives, 85 false negatives and 6 true positives. Some key measures of the model were accuracy = 0.86, AUC = 0.67, TPR = 0.21, FPR = 0.88, MCC = 0.05 and F1 score was 0.10. Additional model performance measures are Table 15.

All relapse models (F1). For the F1 optimization, logistic regression results were the same as accuracy. Evaluating models for AUC, random forest performed the best at (0.67), ridge (0.65), logistic regression (0.63), and LASSO (0.59), but not dramatically different than logistic regression (See Figure 9). Ridge AUC was 6.3 % higher than logistic regression. MCC for ridge was (0.19), logistic regression (0.08) and random forest (0.05). The F1 score for ridge was (0.21), logistic regression (0.12), and random forest was (0.10). Ridge outperformed logistic regression on MCC and F1 score by 122% and 75% respectively. Logistic regression outperformed random forest on MCC and F1 measure. Figure 10 shows the gain curves across

models and upon visual inspection all models need to test about 35-37% of the sample population to capture 50% of positive responses. Additional model performance measures are in Table 22.

Relapse indices. The delta between AUC, MCC, and F1 scores were added and compared for each model pair. This combination of measures was chosen based on model comparison studies in the literature evaluating machine learning model's performance and fit and building upon those studies (Chicco & Jurman, 2020; Couronné et al., 2018; Zhang et al., 2019).

Comparing relapse indices across models' random forest significant outperformed logistic regression and other machine learning algorithms regardless of internal cross-validation optimization. For $\Delta perf_A$ there was a 0.02 or 27.1% difference between random forest and logistic regression for accuracy optimization and $\Delta perf_M$ a 27.5% difference. However, for $\Delta perf_{F,}$ which was the F1 optimization, logistic regression and random forest performed relatively the same (0.82 vs. 0.83). The $\Delta perf_{M1}$ and $\Delta perf_{M2}$ (MCC optimization) LASSO and ridge outperformed logistic regression (0.9%, 9.4%) respectively. The $\Delta perf_{F2}$ index showed that ridge outperformed logistic by 25.8%. The $\Delta perf_{A1}$, $\Delta perf_{A2}$, and $\Delta perf_{F1}$ could not be calculated due to LASSO and ridge models predicted no relapses on all observations, therefore, certain performance measures such as MCC and F1 score could not be calculated. See table 24 and figures 15 and 16 for comprehensive data tabulations and visualizations.

Quantitative Aim 2: All-Cause Utilization Predictive Model Comparison

The second aim of the study is to compare logistic regression to LASSO, ridge and random forest in predicting MS all-cause utilization. MS all-cause utilization included all-cause UC visits plus all-cause ED visits. All-cause UC has low prevalence, for year three there was a total of 258 all-cause UC visits or 10.2% of the total population. The training data set had 181

all-cause UC visits or 10.2% and the test dataset had 77 or 10.2% all-cause UC visits. The low prevalence causes an imbalanced dataset, as there are far many more non-events (no all-cause UC) than events (all-cause UC). An initial backwards logistic regression was fit on the full dataset to set up the training and fit for consistency on all models. Based on the bivariate analysis and full dataset results the following features were included in all relapse models: age, gender, MS center, MS phenotype, DMT, hospitalization, relapse, brain MRI, cervical MRI, thoracic MRI, and all-cause UC as the target variable (outcome). There were 80 models run via cross-validation, for two different optimizations (accuracy, MCC) and 4 different models (logistic, LASSO, ridge, random forest) within each paradigm and eight champion models selected for the endpoint of all-cause UC. There are at total of 80 models and 8 champion models in this aim since the F1 optimization was not included as part of the modeling process. For this endpoint, the F1 results were similar to the other optimization techniques (accuracy, MCC).

Logistic regression (accuracy). A logistic regression using backward elimination was conducted controlling for confounding effects of age, gender and center. Center, MS phenotype, hospitalization, and thoracic MRI were significant at the $p < .05$ level. Relapse visit was fairly significant at $p = .051$. The confusion matrix predicted 647 true negatives, 43 false positives, 35 false negatives and 34 true positives. Some key measures of the model were accuracy = 0.90, AUC = 0.76, TPR = 0.44, FPR = 0.95, MCC = 0.41, and F score = 0.47. Additional model performance measures are in Table 16.

LASSO (accuracy). A LASSO was conducted controlling for confounding effects of age, gender and center. The model's variable importance was as follows: hospitalization (100) the rest of the features had values of 0. The confusion matrix predicted 660 true negatives, 45 false positives, 22 false negatives and 32 true positives. Some key measures of the model were

accuracy = 0.91, AUC = 0.69, TPR = 0.42, FPR = 0.97, MCC = 0.45 and *F* score was 0.49. Additional model performance measures are in Table 17.

Ridge (accuracy). A ridge regression was conducted controlling for confounding effects of age, gender and center. The model's variable importance was as follows: hospitalization (100), PRMS (53.3), RRMS (41.2), center 2 (37.7), other phenotype (36.1), thoracic MRI (33.5), relapse (26.7), center 3 (22.9), SPMS (21.5), center 4 (20.2), cervical MRI (17.4), DMT (2.5), brain MRI (0.67), age (0.49) and gender (0). The confusion matrix predicted 662 true negatives, 59 false positives, 20 false negatives and 18 true positives. Some key measures of the model were accuracy = 0.90, AUC = 0.75, TPR = 0.23, FPR = 0.97, MCC = 0.28 and *F* score was 0.31. Additional model performance measures are in Table 18.

Random forest (accuracy). A random forest was conducted controlling for confounding effects of age, gender and center. The model's variable importance was as follows: hospitalization (100), RRMS (46.0), center 2 (43.4), center 3 (34.0), thoracic MRI (32.6), age (32.1), cervical MRI (28.9), DMT (28.3), center 4 (27.4), relapse (23.7), gender (22.9), other phenotype (21.6), brain MRI (20.3), SPMS (18.8), and PRMS (0). The confusion matrix predicted 665 true negatives, 50 false positives, 17 false negatives and 27 true positives. Some key measures of the model were accuracy = 0.91, AUC = 0.75, TPR = 0.35, FPR = .98, MCC = 0.42 and F1 score was 0.45. Additional model performance measures are in Table 19.

All-UC models (accuracy). Evaluating models for AUC, logistic regression performed the best at (0.76), random forest (0.75), ridge (0.75), and LASSO (0.69) (See Figure 11). Logistic regression's AUC was 1.3% higher than random forest, which was the best performing machine learning model. The MCC for LASSO was (0.45), random forest (0.42) logistic regression (0.41), and ridge (0.28). The F1 score for LASSO was (0.49), logistic regression (0.46), random

forest was (0.44), and ridge (0.31). LASSO outperformed logistic regression on MCC and F1 score by 10% and 6.5% respectively. Figure 12 shows the gain curves across models and upon visual inspection all models need to test about 12.5% of the sample population to capture 50% of positive responses, with the exception of LASSO that requires around 22%. Additional model performance measures are in Table 23.

Ridge (MCC). A ridge regression was conducted controlling for confounding effects of age, gender, and center. The model's variable importance was as follows: hospitalization (100.0), PRMS (50.8), other phenotype (39.9), RRMS (35.0), SPMS (29.9), center 2 (28.6), relapse (25.7), thoracic MRI (23.6) center 4 (18.7), center 3 (15.3), cervical MRI (10.3), brain MRI (3.4), gender (3.3), DMT (2.5), and age (0). The confusion matrix predicted 648 true negatives, 46 false positives, 34 false negatives and 31 true positives. Some key measures of the model were accuracy = 0.90, AUC = 0.76, TPR = 0.40, FPR = 0.95, MCC = 0.38 and F1 score was 0.44. Additional model performance measures are in Table 20.

Random Forest (MCC). A random forest was conducted controlling for confounding effects of age, gender and center. The model's variable importance was as follows: hospitalization (100), RRMS (51.0), center 2 (44.1), thoracic MRI (34.9), center 3 (33.8), DMT (30.0), cervical MRI (30.0), age (29.8), center 4 (26.5), gender (23.2), relapse (22.9), other phenotype (21.9), SPMS (21.7), brain MRI (21.5), and PRMS (0). The confusion matrix predicted 668 true negatives, 50 false positives, 14 false negatives and 27 true positives. Some key measures of the model were accuracy = 0.92, AUC = 0.77, TPR = 0.35, FPR = 0.98, MCC = 0.44 and F1 score was 0.46. Additional model performance measures are in Table 21.

All-UC models (MCC). For the MCC optimization, logistic regression and LASSO results were the same as accuracy. Evaluating models for AUC, all models performed similarly

with the exception of LASSO; random forest (0.76), ridge (0.76), logistic regression (0.76), and LASSO (0.69) (See Figure 13). The MCC for LASSO (0.45), random forest (0.44), logistic regression (0.41), and ridge (0.38). The F1 score for LASSO (0.49), logistic regression (0.47), random forest (0.46), and ridge (0.44). The LASSO outperformed logistic regression on MCC and F1 score by 10% and 4% respectively. Logistic regression outperformed ridge for MCC and random forest and ridge for F1 score. Figure 14 shows the gain curves across models and upon visual inspection all models need to test about 12.5% of the sample population to capture 50% of positive responses, with the exception of LASSO that requires around 22%. Additional model performance measures are in Table 23.

All-UC indices. The delta between AUC, MCC, and F1 score was added and compared for each model pair. When comparing all-cause UC indices across models' logistic regression performed similarly to random forest and LASSO regardless of internal cross-validation optimization. Ridge performed worse overall compared to logistic regression.

For $\Delta perf_A$ there was a -1.1% difference between random forest and logistic regression for accuracy optimization and $\Delta perf_M$ a 1.58% difference for MCC optimization. Random forest and logistic performed similarly on all their indices. The $\Delta perf_{A1}$ index showed that LASSO and logistic regression performed similar (1.631 vs. 1.639), or a -0.5% difference. For $\Delta perf_{A2}$ ridge performed significantly worse over logistic regression (-17.8%). The $\Delta perf_{M2}$ (MCC optimization) ridge performed worse than logistic regression (-3.8%). See table 25 and figures 16 and 17 for comprehensive data tabulations and visualizations.

Chapter 4

Discussion and Conclusions

MS has been studied through basic and clinical trials research and epidemiological studies on population outcomes and quality of life. The MS-CQI adds system-level approaches to study system level effects on population health outcomes using an LHS approach. This study expands upon MS-CQI by moving towards predictive analytics. There is various research using traditional statistical methods, but a lack of research on predictive analytics for MS outcomes. There is a dearth of research on predicting relapse in MS quality improvement collaboratives using predictive analytics, as well as predicting MS relapse and all-cause UC. Relapse and UC utilization are aspects of MS that are costly and disabling, and to date, have been largely unpredictable. Predictive analytics methods have the potential to make these outcomes more predictable than standard approaches and could help shift MS care from a reactive to a predictive approach.

The main purpose of this research is comparing statistical model classes between traditional methods and more sophisticated predictive analytics approaches. There is some literature suggesting what are the "best" model comparison approaches. Zhang et al. (2019) conducted research on a model comparison between random forest and logistic regression in predicting diabetes, and calculated AUC, sensitivity and accuracy for model comparison (2019). Couronné et al. (2018), conducted research and used AIC, AUC, and Brier score and showed the delta between logistic regression and random forest. Couronné et al.'s (2018) study is limited as they only compared a few different classification measures and excluded measures such as F1 Score, and Matthew's correlation coefficient (MCC) which this study will build upon.

In utilizing predictive analytics for MS relapse literature showed some traditional statistical modeling approaches. Research has been used to determine predictors of MS relapse utilizing Poisson regression (Held et al., 2005), predicting RRMS utilizing discrete distribution models to characterize relapsing-remitting CEL (Velez de Mendizabal et al., 2013) and utilizing a multivariate Cox analysis, to predict future relapses in the short term (Sormani et al., 2007). While these statistical approaches are appropriate, other sophisticated predictive modeling approaches could also be evaluated. For Aim 1 of this study, the predictive modeling findings on MS relapse has filled an important gap in literature to compare traditional statistical models to predictive analytics. In addition, the methods used to evaluate the fit and performance of the models have been broadened by creating a more comprehensive index than what is currently in literature that do not include the use of MCC or F1 score (Couronné et al., 2018).

In regards to predicting all-cause UC there is some research on predicting MS hospital utilization. A recent study forecasted MS hospitalization using an Autoregressive Integrated Moving Average (ARIMA; Sharma et al., 2019). There are cost of MS and cost-effectiveness studies estimating current costs and quality of life of patients treated with DMDs (Kobelt et al., 2006). These studies are compelling; however, they do not focus on the endpoint of predicting all-cause UC and do not use predictive analytics and critically evaluate their performance. For Aim 2, this study has filled a significant gap in literature not only with its endpoint of all-cause UC but to compare traditional statistical models to predictive analytics. In addition, the methods used to evaluate the fit and performance of the models are more comprehensive.

Reviewing these numerous models through a data science lens via cross- validation and picking a champion model has made the model more generalizable. Comparing different modeling paradigms to establish optimal predictive approaches for relapses and all-cause UC

could be a trailblazer for the MS community if it is powerful enough to help MS care systems predict, stratify, and address risk in MS populations. All-cause UC has not been researched in depth, various models have not been compared for performance and fit. There are no current uses of modern predictive analytics techniques such as random forest, ridge or LASSO regression for these endpoints in the MS world. This study introduces predictive analytics to the MS field.

Quantitative Discussion Aim 1: Relapse Predictive Model Comparison

Results substantiate that in certain circumstances machine-learning algorithms outperform traditional statistical models. This supports recent popularity of utilizing machine learning algorithm-based approaches such as random forest. Evaluating relapse indices across models, random forest significantly outperformed logistic regression and other machine learning algorithms regardless of internal cross-validation optimization. For the indices calculated in this study, the delta between AUC, MCC, and F1 scores were added and compared for each model pair to create an index.

For $\Delta perf_A$ ($\Delta perf_A = perf_{RF} - perf_{LR}$)there was a 0.23 or 27.1% difference between random forest and logistic regression for accuracy optimization. Decoupling the index, the optimal AUC was 0.69 by the random forest accuracy tuned model, which was 0.06 or 9.5% better than logistic regression (0.63). Couronné et al's (2018) study showed a mean average difference of 0.04 between RF and LR. This study shows an above average difference on that benchmark measure. MCC for random forest was 0.17 versus 0.08 for logistic, a 0.08 or 100% difference. The F1 Score was 0.20 for random forest versus 0.12 for logistic regression, a 0.08 or 69% difference. Random forest outperformed logistic regression for $\Delta perf_A$.

For $\Delta perf_M$ ($\Delta perf_M = perf_{RF} - perf_{LR}$) there was a 0.23 or 27.5% difference between random forest and logistic regression. The AUC on RF was 9.5% higher LR, same as $\Delta perf_A$. In addition,

random forest performed better on MCC and F1 score. These are significant differences when looking at the overall composite index or individual measures. As discussed in the literature MCC is a more appropriate way to evaluate imbalanced datasets (Chicco, & Jurman, 2020). The MMC is not affected by the unbalanced datasets, and is a contingency matrix method of calculating the Pearson product- moment correlation coefficient between actual and predicted values (Chicco, & Jurman, 2020). Thus, the MCC is a reliable statistical measure and produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (Chicco & Jurman, 2020).

For the prediction for $\Delta perf_M$, the true positives were a bit higher in logistic regression (11 vs. 9) compared to random forest. Logistic regression had almost three times the number of false negatives compared to random forest (148 vs. 53) as well as logistic predicting significantly less true negatives (582 vs. 677). The difference between predictive false positives were minimum between RF and LR (20 vs. 18). The number of false negatives in each model drives the overall performance of the logistic model and its associated utility down. Random forest outperformed logistic regression for $\Delta perf_M$.

The imbalanced nature of this dataset limits overall predictive capability, however, RF does a significantly better job at overall prediction across most measures. The increased performance and fit of RF is substantial and does a much better job of predicting true negatives for MS relapse compared to LR. The number of false negatives that LR produces substantiates the use of RF over other modeling techniques when predicting MS relapse.

One of the limitations of the dataset was the imbalance. Relapse is low prevalence and the amount of non-events within the data makes it difficult to make a prediction as some of the statistical models will predict all non-events for relapse. However, as the results have shown,

being able to fine tune models within data science paradigm gives the outcome more flexibility to be accurate. The ability to use different internal cross-validation optimization techniques and oversampling in the data science space allowed further agility and flexibility during the modeling process. In addition, having a wide array of modeling choices beyond traditional statistical modeling allowed greater comparison of different modeling techniques and the opportunity for the "best" model. While the prevalence is low, adding more data may increase the power and robustness of the models.

Quantitative Discussion Aim 2: All-Cause UC Predictive Model Comparison

Results here suggest that machine learning algorithms can perform similarly to traditional statistical models. Here, logistic regression significantly outperformed ridge regression. This finding is an opportunity to better understand the conditions when machine learning algorithms are not preferable to standard statistical methods. When evaluating all-cause UC indices across models' logistic regression performed similarly to random forest and LASSO regardless of internal cross-validation optimization.

For $\Delta perf_A$ there was a -0.02 or -1.1% difference between random forest and logistic regression for accuracy optimization and $\Delta perf_M$ a 0.03 or 1.58% difference. Decoupling the $\Delta perf_M$ index, the optimal AUC was 0.77 by the random forest versus 0.76 for logistic regression a nominal difference. This does not support findings from Couronné et al. (2018) which demonstrated a mean average difference of 4.1% between RF and LR for AUC. This is a below average difference on that benchmark measure. However, MCC for random forest was 0.44 versus 0.41 for logistic, a 0.3 or 7.3% difference. For F1 Score there was minimal difference between RF and LR (.458 vs. .466). For $\Delta perf_A$ random forest vs logistic regression on AUC was similar (0.75 vs. 0.76). In addition, both models performed similar on both MCC and F1 Score.

These are not significant differences when looking at the overall composite index or individual measures. Random forest did not outperform logistic regression on any of the models on this particular measure as well as models optimized for accuracy. There was a negligible difference between the performance of random forest and logistic regression models for $\Delta perf_A$. and $\Delta perf_M$.

For $\Delta perf_{A1}$ ($\Delta perf_{A1}=perf_{LASSO}-perf_{LR}$) comparing LASSO to LR there is a 0.01 (-0.5%) difference on the index. The LR model outperformed LASSO on AUC (0.76 vs. 0.69). However, lasso performed better on the MCC metric (0.45 vs. 0.41). Evaluating LR and LASSO overall via the index there is not much difference between the two models.

For $\Delta perf_{A2}$ ($\Delta perf_{A2}=perf_{RR}-perf_{LR}$) there was a -0.29 or -17.8% difference between the two model indices. The AUC for ridge was 0.751 versus 0.763 for LR, which is not a significant difference. However, when reviewing MCC and F1 score the differences were substantial (0.28 vs. 0.41 and 0.31 vs. 0.46) respectively. This reinforces why an index that includes multiple measures is so pertinent for model fit and performance. Ridge regression had more false positives and less true positives in the prediction (l59 vs. 42 and 18 vs. 34). Logistic regression significantly outperformed ridge regression for $\Delta perf_{A2}$.

The all-cause UC predictive models demonstrate fairly good fit and performance in both random forest and logistic regression. There is not a significant difference between ML algorithms and traditional logistic regression, however, logistic regression does outperform ridge regression. There is an opportunity to better understand in which circumstances ML algorithms may perform better with certain datasets and outcomes such as MS relapse.

Limitations

This study is a secondary analysis based on a Step-Wedge RCT. Randomized clinical trials are the gold standard in research design for inferring a causal relationship (Polit & Beck,

2009). However, there are threats to external validity based on the generalizability of the results. External validity is the extent to which the results of an RCT can be generalized to the general population and in clinical practice (Rothwell, 2005). This analysis was based on year 3 of the study, which is a cross sectional design and is limited to finding associations, and cannot determine causation. The final results of the three-year study which is longitudinal, may be able to thoroughly investigate and confirm the findings presented in this research.

The population observed in the parent MS-CQI study is from a sample of MS care centers in the eastern U.S, and therefor results may have generalizability limited to that geographic region. This geographic limitation impacts the generalizability because it only includes four MS centers located in the Eastern U.S. However, the sample also consists of general demographic characteristics that are representative of PwMS such as majority female and RRMS, which is a strength of the data.

An additional limitation was the inclusion of a limited number of variables from the electronic medical records in the parent MS-CQI study. The inclusion of additional variables that are available within an EHR, such as diagnosis codes and lab results that may have added additional features to the model to increase its predictive power and overall performance.

With both study aims, additional machine learning models could be added and compared such as K Nearest Neighbors, Naïve Bayes and Support Vector Machine. Additional R packages may be available to those with advanced data science expertise that are not utilized as part of this study. Additional data may be advantageous to promote a more robust model for both research aims.

Implications on Future Research

There is an opportunity for future research regarding Aim 1 of this study. Predicting relapse is crucial in MS care. As O'Connell et al. (2014) notated, relapses can result in incomplete recovery and have a significant impact on PwMS quality of life. The ability to apply machine-learning methods such as random forest to predict relapse more accurately could improve the overall quality of life of people with MS as well as save on medical cost. Paired with a learning health system such as MS-CQI which is configured to utilize feed forward predictive analytic data to inform intelligent action, powerful predictive analytics could help convert MS care practice from a reactive to a predictive stance. For example, predictive analytics could help MS care centers identify subpopulations at elevated risk of relapse and take preventive action before the relapses actually occur. In MS-CQI, the majority of patients had RRMS; therefore, there is opportunity to conduct a subsequent independent study of this subpopulation to focused on optimizing relapse outcomes.

There is not much of a cost difference in implementing predictive analytics on smaller scale research. Data science software such as R and Python are open source and are free. If the dataset is small a standard laptop can perform the data wrangling and modeling the cost should be nominal. For smaller scale projects this may be a cost savings as opposed to statistical software that has a license fee associated to use the product. Costs to deploy machine learning on a large scale should be compared to traditional statistical methods to determine utility, cost, and performance. Data scientists are in high demand, costs may be higher for a data scientist versus a traditional statistician. These factors should be evaluated as part of your analytics plan when embarking on new research.

All-cause UC is important in the MS community. The ability to introduce predictive analytics into MS could change health care delivery for MS. Although there was not a big difference in various predictive models, there is a dearth of research on this research aim. This research could improve the overall quality of life of people with MS as well as save on medical cost. In addition, the majority of these patients had RRMS, there is opportunity to conduct a subsequent independent study of this subpopulation to focus on identifying high risk subpopulations and using LHS-oriented improvement and population health approaches to reduce relapse rate.

Conclusion

In this study, logistic regression has been compared to random forest, ridge regression, and LASSO for the first time for feature selection and classification of MS relapse and MS all-cause UC utilization in a population of people with MS followed by four MS centers participating in a LHS model improvement science research collaborative. This study has filled important gaps in literature using a MSCQI based framework. Comparing the predictability of relapse across various models with a predictive analytics framework can potentially change how we manage MS care. Imagine that people with MS could get care and support before a relapse happens because of predictive analytics identifying the patient as high risk, rather than react to it after it happens. This would potentially reduce the risk of disability progression further than could be possible in conventional care.

There are healthcare quality outcomes and cost implications beyond the methodological comparisons of statistical models as theoretical research. Predictive analytics methods have the potential to empower and accelerate the capability of LHS to predict and respond to emerging health needs of people with costly, complex, chronic conditions like MS and better inform

continuous improvement efforts. Predictive analytics could make LHS cheaper, faster, better and could inform research and improvement in new ways and accelerate improvement science research to make MS care better and ultimately, outcomes better for PwMS. This would potentially result in achieving more efficient healthcare outcomes and saving time and overall cost. Sophisticated predictive modeling can make an unpredictable disease like MS more predictable in various outcomes such as relapses and all-cause UC.

References

Backman, D. R., Kohatsu, N. D., Yu, Z., Abbott, R. E., & Kizer, K. W. (2017). Peer Reviewed: Implementing a Quality Improvement Collaborative to Improve Hypertension Control and Advance Million Hearts Among Low-Income Californians, 2014–2015. *Preventing Chronic Disease*, *14*.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, *16*(5), 412-424.

Benneyan, J. C., Lloyd, R. C., & Plsek, P. E. (2003). Statistical process control as a tool for research and healthcare improvement. *BMJ Quality & Safety*, *12*(6), 458-464. http://dx.doi.org/10.1136/qhc.12.6.458

Bozkaya, D., Livingston, T., Migliaccio-Walle, K., & Odom, T. (2017). The cost-effectiveness of disease-modifying therapies for the treatment of relapsing-remitting multiple sclerosis. *Journal of Medical Economics, 20*(3), 297-302. doi:10.1080/13696998.2016.1258366

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199-231. doi:10.1214/ss/1009213726

Cartwright, N. (2010). What are randomised controlled trials good for?. Philosophical studies, 147(1), 59. https://doi.org/10.1007/s11098-009-9450-2

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321-357.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(6), 1-13. https://doi.org/10.1186/s12864-019-6413-7

Clinical microsystems: Where quality, safety and value are made (n.d.). Retrieved from http://www.clinicalmicrosystem.org/

Fanjiang, G., Grossman, J. H., Compton, W. D., & Reid, P. P. (Eds.). (2005). *Building a better delivery system: a new engineering/health care partnership*. National Academies Press. https://doi.org/10.17226/11378

Couronné, R., Probst, P., & Boulesteix, A. L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, *19*(1), 270.

Crohn's & Colitis Foundation. (n.d.). Quality of Care: IBD Qorus. Retrieved from https://www.crohnscolitisfoundation.org/research/ibd-qorus

Definition of MS: National Multiple Sclerosis Society (2019). Retrieved from https://www.nationalmssociety.org/What-is-MS/Definition-of-MS

de Silva, D. (2014). *Improvement collaboratives in health care.* (Report No. 21). London: Health Foundation [PDF file]. Retrieved from https://www.health.org.uk/sites/default/files/ImprovementCollaborativesInHealthcare.pdf

Gisler, S. (2019, May 21). EMD Serono Launches MS-LINK Research Network to improve patient care. Retrieved from https://multiplesclerosisnewstoday.com/news-posts/2019/05/07/emd-serono-launches-ms-link-a-community-to-advance-ms-science-and-improve-patient-lives/

Godfrey, M. M., & Oliver, B. J. (2014). Accelerating the rate of improvement in cystic fibrosis care: Contributions and insights of the learning and leadership collaborative. *BMJ Quality & Safety*, *23*(Suppl 1), i23-i32. doi:10.1136/bmjqs-2014-002804

Gustafson, D. H., Quanbeck, A. R., Robinson, J. M., Ford, J. H., Pulvermacher, A., French, M. T.,...McCarty, D. (2013). Which elements of improvement collaboratives are most effective? A cluster-randomized trial. *Addiction*, *108*(6), 1145-1157. doi:10.1111/add.12117

Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian Journal of Internal Medicine*, *4*(2), 627-635.

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377-381. https://doi.org/10.1016/j.jbi.2008.08.010

Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L.,...Duda, S. N. (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics*, *95*, 103208. https://doi.org/10.1016/j.jbi.2019.103208

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* (2nd ed.). New York: Springer.

Held, U., Heigenhauser, L., Shang, C., Kappos, L., & Polman, C. (2005). Predictors of relapse rate in MS clinical trials. *Neurology, 65*(11), 1769-773. doi:10.1212/01.wnl.0000187122.71735.1f

Hemming, K., Haines, T. P., Chilton, P. J., Girling, A. J., & Lilford, R. J. (2015). The stepped wedge cluster randomised trial: Rationale, design, analysis, and reporting. *BMJ*, *350*, 1-7. doi:10.1136/bmj.h391

Hulscher. M., Schouten, L., & Grol, R. (2009). C*ollaboratives*. London: Health Foundation [PDF file]. Retrieved from https://www.health.org.uk/sites/default/files/Collaboratives.pdf

Iezzoni, L. I., & Ngo, L. (2007). Health, disability, and life insurance experiences of working-age persons with multiple sclerosis. *Multiple Sclerosis Journal*, *13*(4), 534-546

Institute for healthcare improvement: The IHI triple aim. (n.d.). Retrieved from from http://www.ihi.org:80/Engage/Initiatives/TripleAim/Pages/default.aspx

Institute for Healthcare Improvement (2014). Plan-Do-Study-Act (PDSA) worksheet. Retrieved from http://www.ihi.org/knowledge/Pages/Tools/PlanDoStudyActWorksheet.aspx

Institute for Healthcare Improvement (2003). *The Breakthrough Series: IHI's Collaborative Model for Achieving Breakthrough Improvement* [PDF file]. Retrieved from http://www.ihi.org/resources/Pages/IHIWhitePapers/TheBreakthroughSeriesIHIsCollaborativeModelforAchievingBreakthroughImprovement.aspx

Institute of Medicine (2001). *Crossing the quality chasm: A new health system for the 21st century* [PDF file]. Retrieved from http://www.iom.edu/~/media/Files/Report%20Files/2001/Crossing-the-Quality-Chasm/Quality%20Chasm%202001%20%20report%20brief.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York, NY: Springer.

Jones, E., Pike, J., Marshall, T., & Ye, X. (2016). Quantifying the relationship between increased disability and health care resource utilization, quality of life, work productivity, health

care costs in patients with multiple sclerosis in the US. *BMC Health Services Research*, *16*(1), 294-302. doi:10.1186/s12913-016-1532-1

Kilo, C. M. (1998). A framework for collaborative improvement: Lessons from the Institute for Healthcare Improvement's Breakthrough Series. *Quality Management in Health Care*, *6*, 1-14.

Kirasich, K., Smith, T., & Sadler, B. (2018). Random forest vs logistic regression: Binary classification for heterogeneous datasets. *SMU Data Science Review*, *1*(3), 1-24. Retrieved from https://scholar.smu.edu/datasciencereview/vol1/iss3/9

Kobelt, G., Berg, J., Atherly, D., & Hadjimichael, O. (2006). Costs and quality of life in multiple sclerosis. *Neurology, 66*(11), 1696-702. doi:10.1212/01.wnl.0000218309.01322.5c

Koch-Henriksen, N., Thygesen, L. C., Sørensen, P. S., & Magyari, M. (2019). Worsening of disability caused by relapses in multiple sclerosis: A different approach. *Multiple sclerosis and related disorders*, *32*, 1-8.

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development* (Vol. 1). Englewood Cliffs, NJ: Prentice-Hall.

Lean Six Sigma Online. (n.d.). What are the Lean Six Sigma principles? Retrieved from https://www.purdue.edu/leansixsigmaonline/blog/lean-six-sigma-principles/

Lindblad, S., Ernestam, S., Van Citters, A. D., Lind, C., Morgan, T. S., & Nelson, E. C. (2017). Creating a culture of health: Evolving healthcare systems and patient engagement. *QJM: An International Journal of Medicine*, *110*(3), 125-129. https://doi.org/10.1093/qjmed/hcw188

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, *405*(2), 442-451. https://doi.org/10.1016/0005-2795(75)90109-9

Meninno, E., Livingston, T., Stuart, S., Powell, S., Ahmad, A., Afsari, N., ... & Tornatore, C. (2018). Observations of a multiple sclerosis patient-centered specialty practice: Analysis of depression patterns. *Neurology Apr 2018, 90* (15 Supplement) P4. 425. Retrieved from https://n.neurology.org/content/90/15_Supplement/P4.425.abstract

Multiple Sclerosis Partners Advancing Technology and Health Solutions. (2016). Seeking to enhance MS care. Retrieved from https://www.mspaths.com

Nelson, E. C., Greenfield, S., Hays, R. D., Larson, C., Leopold, B., & Batalden, P. B. (1995). Comparing outcomes and charges for patients with acute myocardial infarction in three community hospitals: An approach for assessing "value". *International Journal for Quality in Health Care, 7*(2), 95-108. doi:10.1093/intqhc/7.2.95

Nelson, E. C., Batalden, P. B., Godfrey, M. M., & Lazar, J. S. (2011). *Value by design: developing clinical microsystems to achieve organizational excellence*. John Wiley & Sons

Nelson, E. C., Dixon-Woods, M., Batalden, P. B., Homa, K., Van Citters, A. D., Morgan, T. S.,...Lindblad, S. (2016). Patient focused registries can improve health, care, and science. *BMJ, 354*, 1-6. doi:10.1136/bmj.i3319

North American Research Committee on Multiple Sclerosis Sclerosis, (2017). NARCOMS registry for multiple sclerosis. Retrieved from https://www.narcoms.org/

O'Connell, K., Kelly, S. B., Fogarty, E., Duggan, M., Buckley, L., Hutchinson, M.,...Tubridy, N. (2014). Economic costs associated with an MS relapse. *Multiple Sclerosis and Related Disorders, 3*(6), 678-683. doi:10.1016/j.msard.2014.09.002

Oliver, B. J. (2019, October, 2019). The Multiple Sclerosis Continuous Improvement Collaborative: A coproduction research learning health system for costly complex chronic illness care. Platform presentation at The International Society for Quality in Health Care (ISQua) Annual International Conference. Cape Town, South Africa

Oliver, B. J., Messier, R., Hall, A. for the MSCQI Investigators, (2019, June 2019). The Multiple Sclerosis Continuous Improvement Collaborative: Improvement science and coaching applications in a multicenter outcomes research study. Paper presented at the Sheffield Microsystem Coaching Academy Annual International Expo. Sheffield, United Kingdom

Polit, D. F., & Beck, C. T. (2009). International differences in nursing research, 2005–2006. *Journal of Nursing Scholarship*, *41*(1), 44-53.

Porter, M. E. (2009). A strategy for health care reform—toward a value-based system. *New England Journal of Medicine*, *361*(2), 109-112.

Quality of Care: IBD Qorus (n.d.). Retrieved from https://www.crohnscolitisfoundation.org/research/ibd-qorus

Quinn, J. B. (1992). *Intelligent enterprise: A knowledge and service based paradigm for industry*. Free Press, New York, USA.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Riccio, P., Rossano, R., Larocca, M., Trotta, V., Mennella, I., Vitaglione, P., ...Coniglio, M. G. (2016). Anti-inflammatory nutritional intervention in patients with relapsing-remitting and primary-progressive multiple sclerosis: A pilot study. *Experimental Biology and Medicine*, *241*(6), 620-635. https://doi.org/10.1177/1535370215618462

Sharma, K., Bittner, F., & Cho, T. (2019). Forecasting multiple sclerosis hospitalization and healthcare burden in the United States from 2017 to 2040. *Neurology*, *92*(Suppl 15), P4. 6-007.

Šimundić A. M. (2009). Measures of diagnostic accuracy: Basic definitions. *eJournal of the International Federation of Clinical Chemistry and Laboratory Medicine, 19*(4), 203–211.

Sormani, M. P., Rovaris, M., Comi, G., & Filippi, M. (2007). A composite score to predict short-term disease activity in patients with relapsing-remitting MS. *Neurology, 69*(12), 1230-1235. doi:10.1212/01.wnl.0000276940.90309.15

Stoppe, M., Busch, M., Krizek, L., & Then Bergh, F. (2017). Outcome of MS relapses in the era of disease-modifying therapy. *BMC Neurology, 17*(1), 151-159. doi:10.1186/s12883-017-0927-x

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2019). *Using multivariate statistics* (7th ed.). Boston, MA: Pearson.

Thor, J., Lundberg, J., Ask, J., Olsson, J., Carli, C., Härenstam, K. P., & Brommels, M. (2007). Application of statistical process control in healthcare improvement: Systematic review. *BMJ Quality & Safety*, *16*(5), 387-399. doi:10.1136/qshc.2006.022194

Types of MS: National Multiple Sclerosis Society (2019). Retrieved from https://www.nationalmssociety.org/What-is-MS/Types-of-MS

Velez de Mendizabal, N., Hutmacher, M. M., Troconiz, I. F., Goni, J., Villoslada, P., Bagnato, F., & Bies, R. R. (2013). Predicting relapsing-remitting dynamics in multiple sclerosis using discrete distribution models: A population approach. *PloS One, 8*(9), 1-11. doi:10.1371/journal.pone.0073361

Vermont Oxford Network: Who We are. (n.d.). Retrieved from https://public.vtoxford.org/who-we-are-overview/

Vollmer, T. (2007). The natural history of relapses in multiple sclerosis. *Journal of the Neurological Sciences, 256*, S5-S13. https://doi.org/10.1016/j.jns.2007.01.065

Wallin, M. T., Culpepper, W. J., Campbell, J. D., Nelson, L. M., Langer-Gould, A., Marrie, R. A.,...LaRocca, N. G. (2019). The prevalence of MS in the United States: A population-based estimate using health claims data. *Neurology*, *92*(10), e1029-e1040. doi:10.1212/wnl.0000000000007035

West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D.,...Mullen, P. D. (2008). Alternatives to the randomized controlled trial. *American Journal of Public Health*, *98*(8), 1359-1366. doi:10.2105/AJPH.2007.124446

What are the Lean Six Sigma Principles? (n.d.). Retrieved from https://www.purdue.edu/leansixsigmaonline/blog/lean-six-sigma-principles/

Zhang, B., Lu, L., & Hou, J. (2019). A comparison of logistic regression, random forest models in predicting the risk of diabetes. *ISICDM 2019: Proceedings of the Third International Symposium on Image Computing and Digital Medicine*. 231-234. https://doi.org/10.1145/3364836.3364882.

Table 1

Patient Characteristics for Year 3 by Center

	Center 1		Center 2		Center 3		Center 4		Total	
Characteristics	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Participants	977	38.6	460	18.2	539	21.3	556	22.0	2,532	100
Phenotype***										
RRMS	787	80.6	364	79.1	393	72.9	518	93.2	2,062	81.4
SPMS	89	9.1	43	9.3	78	14.5	18	3.2	228	9.0
PPMS	59	6.0	28	6.1	37	6.9	12	2.2	136	5.4
PRMS	0	0.0	0	0.0	0	0.0	7	1.3	7	0.3
Other	42	4.3	25	5.4	31	5.8	1	0.18	99	3.9
Age (mean, SD) ***	51.3	12.2	48.3	13.2	50.6	12.8	48.0	12.9	49.9	12.7
Sex***										
Female	764	78.2	332	72.2	384	71.2	437	78.6	1,917	75.7
Male	213	22.0	128	27.9	155	28.8	119	21.4	615	24.3

Note. Chi square for categorical variables, ANOVA for continuous variables. * p<.05, ** p<.01, *** p<.001.

Table 2

Outcomes for Year 3 by Center

	Center 1		Center 2		Center 3		Center 4		Total	
Outcome	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Participants	977	38.6	460	18.2	539	21.3	556	22.0	2,532	100
>=1DMT***	719	73.6	274	59.6	327	60.7	526	94.6	1,846	72.9
Brain MRI	498	51.0	209	45.4	199	36.9	311	55.9	1,217	48.1
Cervical MRI***	338	34.6	112	24.4	99	18.4	129	23.2	678	26.8
Thoracic MRI***	142	14.5	106	23.0	21	3.9	101	18.2	370	14.6
Hospitalizations***	57	5.8	49	10.7	23	4.3	47	8.5	176	7.0
ER Visits***	79	8.1	51	11.1	31	5.8	46	8.3	207	8.2
UC Visits***	34	3.5	24	5.2	9	1.7	1	0.2	68	2.7
Relapses***	33	3.4	14	3.0	14	2.6	36	6.5	97	3.8
Hospitalization (mean, SD)**	0.07	0.31	0.15	0.48	0.05	0.23	0.09	0.32	0.09	0.34
ED Utilization (mean, SD)**	0.10	0.36	0.15	0.45	0.06	0.28	0.09	0.30	0.1	0.35
UC Utilization (mean, SD)**	0.04	0.21	0.06	0.27	0.02	0.13	0.002	0.04	0.03	0.19
Relapse Rate (mean, SD)**	0.04	0.20	0.04	0.23	0.03	0.18	0.06	0.27	0.04	0.22

Note. Chi square for categorical variables, ANOVA for continuous variables. * p<.05, ** p<.01, *** p<.001.

Table 3

Patient Characteristics and Outcomes by Derivation and Validation Cohort

	Total		Derivation Cohort		Validation Cohort	
Characteristic	Number	Percent	Number	Percent	Number	Percent
Participants	2,532	100	1,773	70.0	759	30.0
Age (mean, SD)	49.9	12.7	49.9	12.8	49.9	12.6
Sex						
Male	615	24.2	434	24.4	181	13.5
Female	1,917	75.8	1,339	75.6	578	86.5
Phenotype						
RRMS						
No	470	18.5	313	17.6	157	20.1
Yes	2,062	81.5	1,460	82.4	602	79.9
SPMS						
No	2,304	90.9	1,619	92.3	685	90.2
Yes	228	9.1	154	7.7	74	9.8
PPMS						
No	2,396	94.6	1,684	95.0	712	93.8
Yes	136	5.4	89	5.0	47	6.2

	Total		Derivation Cohort		Validation Cohort	
Characteristic	Number	Percent	Number	Percent	Number	Percent
PRMS						
No	2,525	99.7	1,770	99.8	755	99.5
Yes	7	0.0	3	0.0	4	0.0
Other						
No	2,433	96.0	1,706	96.2	727	95.8
Yes	99	4.0	67	3.8	32	4.2
DMT						
No	686	27.1	468	26.4	218	28.7
Yes	1,846	72.9	1,305	73.6	541	71.3
Brain MRI						
No	1,315	51.9	901	50.8	414	54.5
Yes	1,217	48.1	872	49.2	345	45.5
Cervical MRI						
No	1,854	73.2	1,283	72.3	571	75.2
Yes	678	26.8	490	27.7	188	24.3
All UC						
No	2,274	89.8	1,592	89.8	682	89.8

	Total		Derivation Cohort		Validation Cohort	
Characteristic	Number	Percent	Number	Percent	Number	Percent
Yes	258	10.2	181	10.2	77	10.2
Thoracic MRI						
No	2,162	85.4	1,506	86.4	656	86.4
Yes	370	14.6	267	13.6	103	13.6
Hospitalizations						
No	2,356	93.0	1,651	93.1	705	92.9
Yes	176	7.0	122	6.9	54	7.1
ED Visits						
No	2,325	91.8	1,628	91.8	697	91.8
Yes	207	8.2	145	8.2	62	8.2
UC Visits						
No	2,464	97.3	1,725	97.3	739	97.4
Yes	68	2.7	48	2.7	20	2.6
Relapse						
No	2,435	96.2	1,699	95.8	736	96.9
Yes	97	3.8	74	4.2	23	3.1

Table 4

Bivariate Analysis (n = 2,532)

Variable	1	2	3	4	5	6	7	8	9	10	11	12
1. Phenotype	1	¥	30.0***	164.2***	30.3***	11.3*	10.5*	16.6*	4.1	1.06	7.6	122.1***
2. Age	¥	1	¥	¥	¥	¥	¥	¥	¥	¥	¥	¥
3. Gender	30.0***	¥	1	1.0	1.32	0.06	0.81	0.1	.40	4.6*	0.38	14.8*
4. DMT	164.2***	¥	1.0	1	177.2***	64.8***	45.4***	2.1	.10	0.45	8.2**	215***
5. Brain MRI	30.3***	¥	1.3	177.2***	1	781.1***	329.3***	2.2	5.7*	0.67	29.9***	45***
6. Cervical MRI	11.3*	¥	.06	64.8***	781.1***	1	944.8***	1.47	9.3**	3.6	31.6***	54.9***
7. Thoracic MRI	10.5*	¥	.81	45.4***	329.3***	944.8***	1	6.2*	11.8**	6.0*	33.8***	81.4***
8. Hospitalization	16.6**	¥	.10	2.1	2.2	1.47	6.2*	1	582.3***	20.1***	11.3**	19.6***
9. ED Visit	4.1	¥	.40	0.1	5.7*	9.25**	11.8**	582.3**	1	26.3***	17.5***	9.4*
10. UC Visit	1.06	¥	4.6*	0.45	.67	3.55	6.0*	20.1***	26.3***	1	0.06	29.1***
11. Relapse	7.6	¥	.38	8.2**	29.9***	31.6***	33.8***	11.3**	17.5***	0.06	1	14.1**
12. Center	122***	¥	14.8**	215.1***	45.2***	54.9***	81.4***	19.6**	9.4*	29.1***	14.1**	1

Note. Chi square for categorical variables, Pearson *r* for continuous variables. ED=emergency department; UC=urgent care. * p <0.05, ** p<0.01, *** p<0.001. Test could not be performed = ¥.

Table 5

Model Comparison Metrics

Metric	Formula
Accuracy	(true positives + true negatives) / (true positives + true negatives + false positives + false negatives)
True Positive Rate (TPR)-Sensitivity	true positives / (true positives + false negatives)
False Positive Rate (FPR)-Specificity	false positives / (false positives + true negatives)
Negative Predictive Value (NPV)	True negatives / (true negatives + false negatives)
Precision (PPV)	true positives / (true positives + false positives)
Area Under the Curve (AUC)	Graphical Plotting of TPR vs. FPR
Matthews Correlation Coefficient (MCC)	$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$

Note. Adapted from "Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets," by K. Kirasich, T. Smith and B. Sadler, 2018, *SMU Data Science Review, 1*(3), p. 13.

Table 6

Relapse Model Comparison Metrics-Cross-Validated (Accuracy) Stepwise Logistic Regression

Metric	Result
Accuracy	0.781
True Positive Rate (TPR)	0.379
False Positive Rate (FPR)	0.797
Negative Predictive Value (NPV)	0.970
MCC	0.083
Precision	0.069
F Score	0.117
AUC	0.629

Table 7

Relapse Model Comparison Metrics-Cross-Validated (Accuracy) LASSO

Metric	Result
Accuracy	0.962
True Positive Rate (TPR)	0.000
False Positive Rate (FPR)	1.000
Negative Predictive Value (NPV)	0.962
MCC	¥
Precision	¥
F Score	¥
AUC	0.510

Note. Test could not be performed = ¥.

Table 8

Relapse Model Comparison Metrics-Cross-Validated (Accuracy) Ridge

Metric	Result
Accuracy	0.962
True Positive Rate (TPR)	0.000
False Positive Rate (FPR)	1.000
Negative Predictive Value (NPV)	0.962
MCC	¥
Precision	¥
F Score	¥
AUC	0.670

Note. Test could not be performed = ¥.

Table 9

Relapse Model Comparison Metrics-Cross-Validated (Accuracy) Random Forest

Metric	Result
Accuracy	0.904
True Positive Rate (TPR)	0.310
False Positive Rate (FPR)	0.927
Negative Predictive Value (NPV)	0.971
MCC	0.166
Precision	0.145
F Score	0.198
AUC	0.690

Table 10

Relapse Model Comparison Metrics-Cross-Validated (MCC) LASSO

Metric	Result
Accuracy	0.773
True Positive Rate (TPR)	0.414
False Positive Rate (FPR)	0.788
Negative Predictive Value (NPV)	0.971
MCC	0.093
Precision	0.071
F Score	0.122
AUC	0.622

Table 11

Relapse Model Comparison Metrics-Cross-Validated (MCC) Ridge

Metric	Result
Accuracy	0.810
True Positive Rate (TPR)	0.414
False Positive Rate (FPR)	0.826
Negative Predictive Value (NPV)	0.973
MCC	0.119
Precision	0.086
F Score	0.143
AUC	0.645

Table 12

Relapse Model Comparison Metrics-Cross-Validated (MCC) Random Forest

Metric	Result
Accuracy	0.895
True Positive Rate (TPR)	0.345
False Positive Rate (FPR)	0.916
Negative Predictive Value (NPV)	0.972
MCC	0.172
Precision	0.141
F Score	0.200
AUC	0.685

Table 13

Relapse Model Comparison Metrics-Cross-Validated (F1) LASSO

Metric	Result
Accuracy	0.962
True Positive Rate (TPR)	0.000
False Positive Rate (FPR)	0.000
Negative Predictive Value (NPV)	0.962
MCC	¥
Precision	¥
F Score	¥
AUC	0.629

Note. Test could not be performed = ¥.

Table 14

Relapse Model Comparison Metrics-Cross-Validated

Metric	Result
Accuracy	0.877
True Positive Rate (TPR)	0.414
False Positive Rate (FPR)	0.896
Negative Predictive Value (NPV)	0.975
MCC	0.185
Precision	0.136
F Score	0.205
AUC	0.653

Table 15

Relapse Model Comparison Metrics-Cross-Validated (F1) Random Forest

Metric	Result
Accuracy	0.858
True Positive Rate (TPR)	0.207
False Positive Rate (FPR)	0.884
Negative Predictive Value (NPV)	0.966
MCC	0.053
Precision	0.065
F Score	0.10
AUC	0.667

Table 16

All-Cause Urgent Care Model Comparison Metrics-Cross-Validated (Accuracy) Stepwise Logistic Regression

Metric	Result
Accuracy	0.897
True Positive Rate (TPR)	0.442
False Positive Rate (FPR)	0.949
Negative Predictive Value (NPV)	0.938
MCC	0.410
Precision	0.493
F Score	0.466
AUC	0.763

Table 17

All-Cause Urgent Care Model Comparison Metrics-Cross-Validated (Accuracy) LASSO

Metric	Result
Accuracy	0.912
True Positive Rate (TPR)	0.416
False Positive Rate (FPR)	0.968
Negative Predictive Value (NPV)	0.936
MCC	0.450
Precision	0.593
F Score	0.489
AUC	0.692

Table 18

All-Cause Urgent Care Model Comparison Metrics-Cross-Validated (Accuracy) Ridge

Metric	Result
Accuracy	0.896
True Positive Rate (TPR)	0.234
False Positive Rate (FPR)	0.971
Negative Predictive Value (NPV)	0.918
MCC	0.283
Precision	0.474
F Score	0.313
AUC	0.751

Table 19

All-Cause Urgent Care Model Comparison Metrics-Cross-Validated (Accuracy) Random Forest

Metric	Result
Accuracy	0.912
True Positive Rate (TPR)	0.351
False Positive Rate (FPR)	0.975
Negative Predictive Value (NPV)	0.930
MCC	0.421
Precision	0.614
F Score	0.446
AUC	0.754

Table 20

All-Cause Urgent Care Model Comparison Metrics-Cross-Validated (MCC) Ridge

Metric	Result
Accuracy	0.895
True Positive Rate (TPR)	0.403
False Positive Rate (FPR)	0.950
Negative Predictive Value (NPV)	0.934
MCC	0.381
Precision	0.477
F Score	0.437
AUC	0.758

Table 21

All-Cause Urgent Care Model Comparison Metrics-Cross-Validated (MCC) Random Forest

Metric	Result
Accuracy	0.916
True Positive Rate (TPR)	0.351
False Positive Rate (FPR)	0.979
Negative Predictive Value (NPV)	0.930
MCC	0.441
Precision	0.659
F Score	0.458
AUC	0.766

Table 22

Relapse Model Comparison Metrics-All Predictive Models

Model	Accuracy	TPR	FPR	NPV	MCC	Precision	F1 Score	AUC
Logistic Regression (Accuracy)	0.781	0.379	0.797	0.97	0.083	0.069	0.117	0.629
LASSO (Accuracy)	0.962	0	1	0.962	¥	¥	¥	0.51
Ridge (Accuracy)	0.962	0	1	0.962	¥	¥	¥	0.67
Random Forest (Accuracy)	0.904	0.310	0.927	0.971	0.166	0.145	0.198	0.69
Logistic Regression (MCC)	0.781*	0.379*	0.797*	0.97*	0.083*	0.069*	0.117*	0.629*
LASSO (MCC)	0.773	0.414	0.788	0.971	0.093	0.071	0.122	0.622
Ridge (MCC)	0.810	0.415	0.826	0.973	0.119	0.086	0.143	0.645
Random Forest (MCC)	0.895	0.345	0.916	0.972	0.172	0.141	0.2	0.685
Logistic Regression (F1)	0.781*	0.379*	0.797*	0.97*	0.083*	0.069*	0.117*	0.629*
LASSO (F1)	0.962	0	1	0.962	¥	¥	¥	0.585
Ridge (F1)	0.877	0.414	0.896	0.975	0.185	0.136	0.205	0.653
Random Forest (F1)	0.858	0.207	0.884	0.966	0.053	0.065	0.10	0.667

Note. * Logistic regression results are the same across all optimization techniques. Test could not be performed = ¥.

Table 23

All-Cause Urgent Care Model Comparison Metrics-All Predictive Models

Model	Accuracy	TPR	FPR	NPV	MCC	Precision	F1 Score	AUC
Logistic Regression (Accuracy)	0.897	0.442	0.949	0.938	0.410	0.493	0.466	0.763
LASSO (Accuracy)	0.912	0.416	0.968	0.936	0.450	0.593	0.489	0.692
Ridge (Accuracy)	0.896	0.234	0.971	0.918	0.283	0.474	0.313	0.751
Random Forest (Accuracy)	0.912	0.351	0.975	0.930	0.421	0.614	0.446	0.754
Logistic Regression (MCC)	0.897*	0.442*	0.949*	0.938*	0.410*	0.493*	0.466*	0.763*
LASSO (MCC)	0.912*	0.416*	0.968*	0.936*	0.450*	0.593*	0.489*	0.692*
Ridge (MCC)	0.895	0.403	0.950	0.934	0.381	0.477	0.437	0.758
Random Forest (MCC)	0.916	0.351	0.979	0.930	0.441	0.659	0.458	0.766

Note. * Logistic regression results are the same across all optimization techniques.

Table 24

Performance Relapse Indices

Index	Result
$\Delta perf_A = perf_{RF} - perf_{LR}$	1.054-.829=0.225 (27.1%)
$\Delta perf_{A1} = perf_{LASSO} - perf_{LR}$	-¥ (MCC and F1 not calculated for LASSO)
$\Delta perf_{A2} = perf_{RR} - perf_{LR}$	-¥ (MCC and F1 not calculated for Ridge)
$\Delta perf_M = perf_{RF} - perf_{LR}$	1.057-.829*=0.228 (27.5%)
$\Delta perf_{M1} = perf_{LASSO} - perf_{LR}$	0.837-0.829*=0.008 (0.9%)
$\Delta perf_{M2} = perf_{RR} - perf_{LR}$	.907-.829*=0.078 (9.4%)
$\Delta perf_F = perf_{RF} - perf_{LR}$	0.820-0.829*=-0.009 (-1.1%)
$\Delta perf_{F1} = perf_{LASSO} - perf_{LR}$	-¥ (MCC and F1 not calculated) for LASSO
$\Delta perf_{F2} = perf_{RR} - perf_{LR}$	1.043-.829*=0.214 (25.8%)

Note. Results are same as accuracy optimization = *. MCC and F1Test could not be performed = ¥.

Table 25

Performance All-Cause Urgent Care Indices

Index	Result
$\Delta perf_A = perf_{RF} - perf_{LR}$	1.621-1.639=-0.018 (-1.1%)
$\Delta perf_{A1} = perf_{LASSO} - perf_{LR}$	1.631-1.639-0.008 (-0.5%)
$\Delta perf_{A2} = perf_{RR} - perf_{LR}$	1.347-1.639=0.292 (-17.8%)
$\Delta perf_M = perf_{RF} - perf_{LR}$	1.665-1.639=0.026 (1.58%)
$\Delta perf_{M1} = perf_{LASSO} - perf_{LR}$	*
$\Delta perf_{M2} = perf_{RR} - perf_{LR}$	1.665-1.639=-0.063 (-3.84%)

Note. Results are same as accuracy optimization= *.

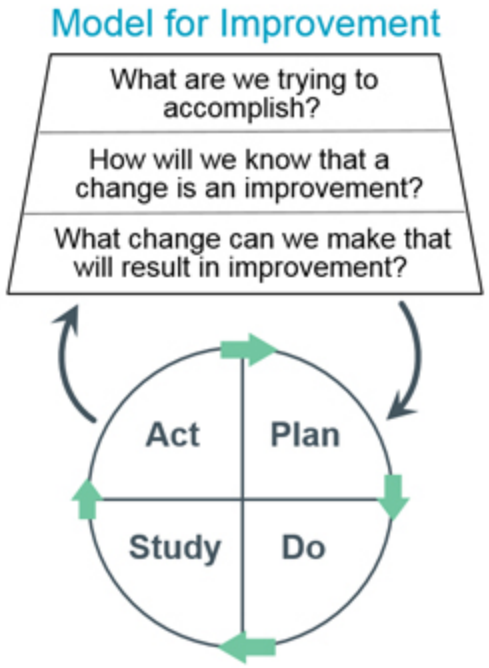


Figure 1. A Diagram of IHI Model for Improvement Showing Plan-Do-Study-Act process. Taken from "How to Improve," by Institute for Healthcare Improvement, n.d. (http://www.ihi.org/resources/Pages/HowtoImprove/default.aspx). Copyright 2019 by John Wiley and Sons.

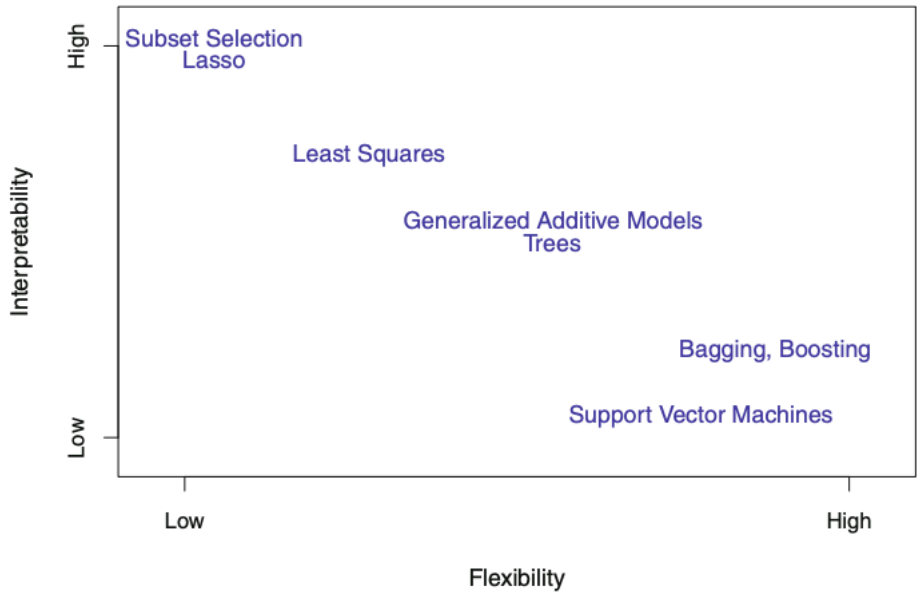


Figure 2. Representation of the Tradeoff Between Flexibility and Interpretability Using Different Statistical Learning Methods. Taken from *Introduction to Statistical Learning: With Applications in R* (p. 25), by G. James, D. Witten, T. Hastie, and R. Tibshirani, 2013, New York, NY: Springer Nature. Copyright 2013 by Springer Nature.

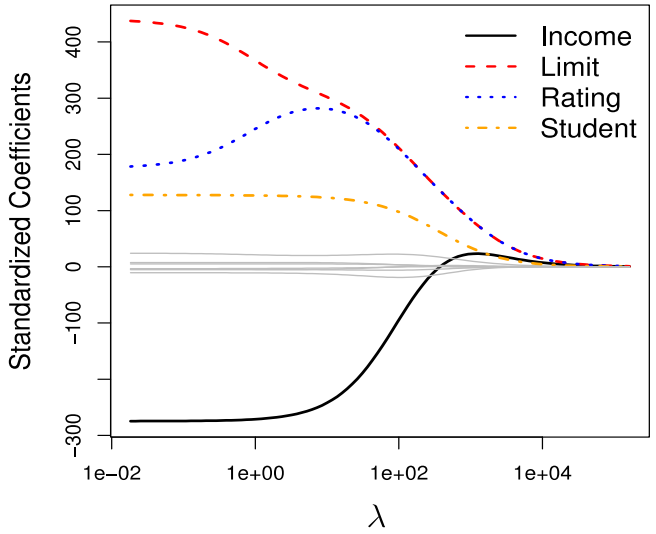


Figure 3. Ridge Regression Coefficient Estimates for Each Value of λ. Taken from *Introduction to Statistical Learning: With Applications in R* (p. 216), by G. James, D. Witten, T. Hastie, and R. Tibshirani, 2013, New York, NY: Springer Nature. Copyright 2013 by Springer Nature.

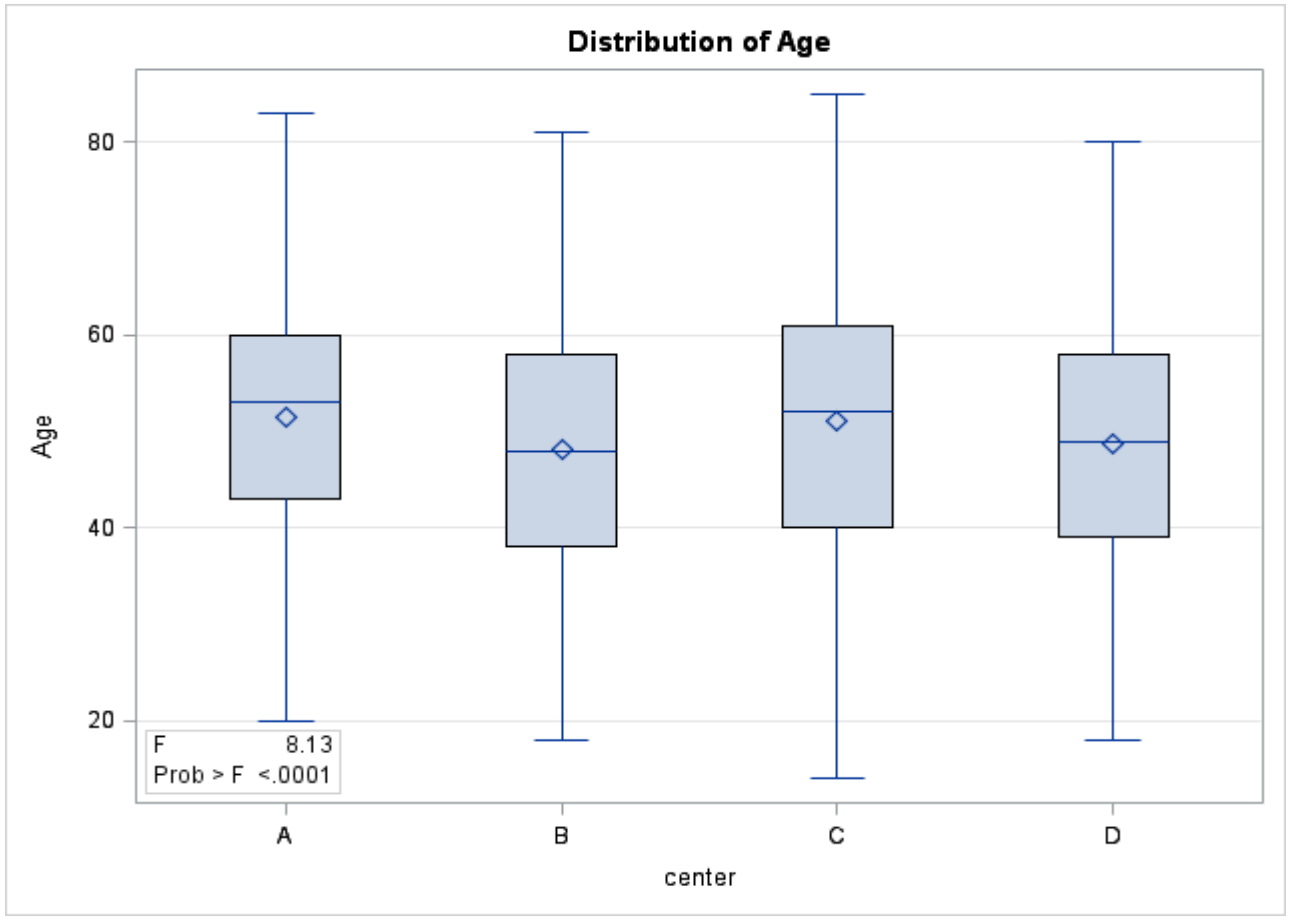


Figure 4. Box Plot of Age by Center. Copyright 2020 by SAS 9.4.

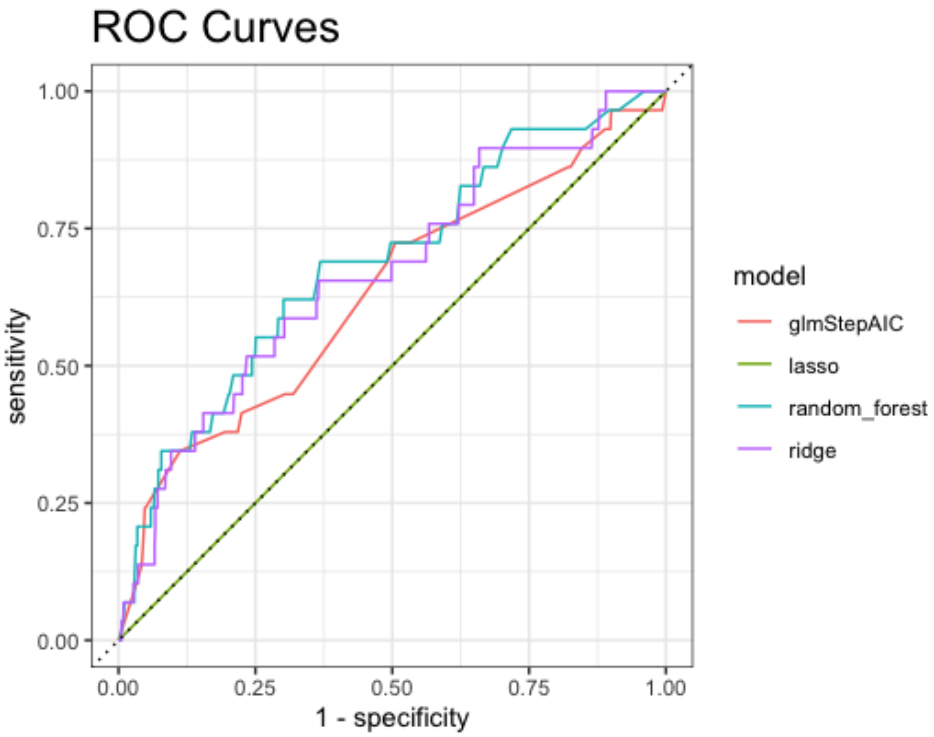


Figure 5. Relapse ROC Curve Model Comparison Optimized for Accuracy. Copyright 2020 by R.

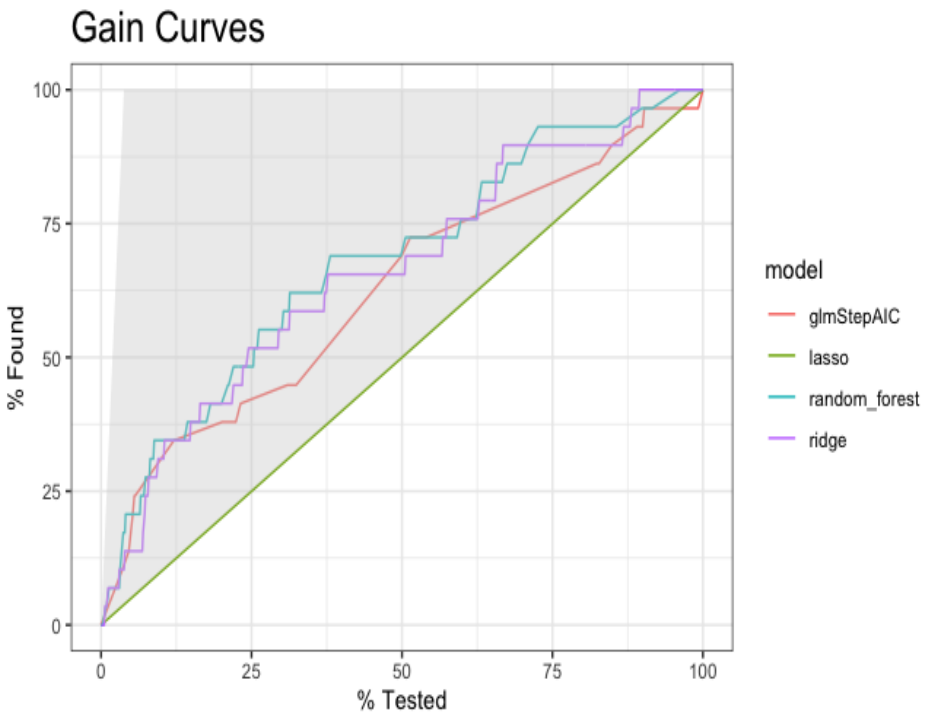


Figure 6. Relapse Gain Curve Model Comparison Optimized for Accuracy. Copyright 2020 by R.

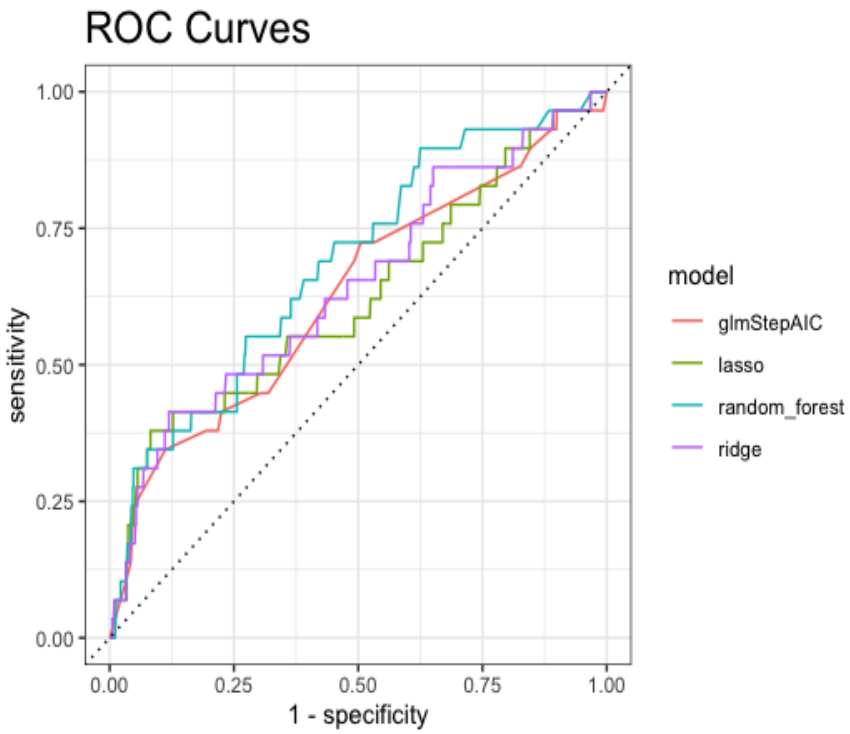


Figure 7. Relapse ROC Curve Model Comparison Optimized for MCC.

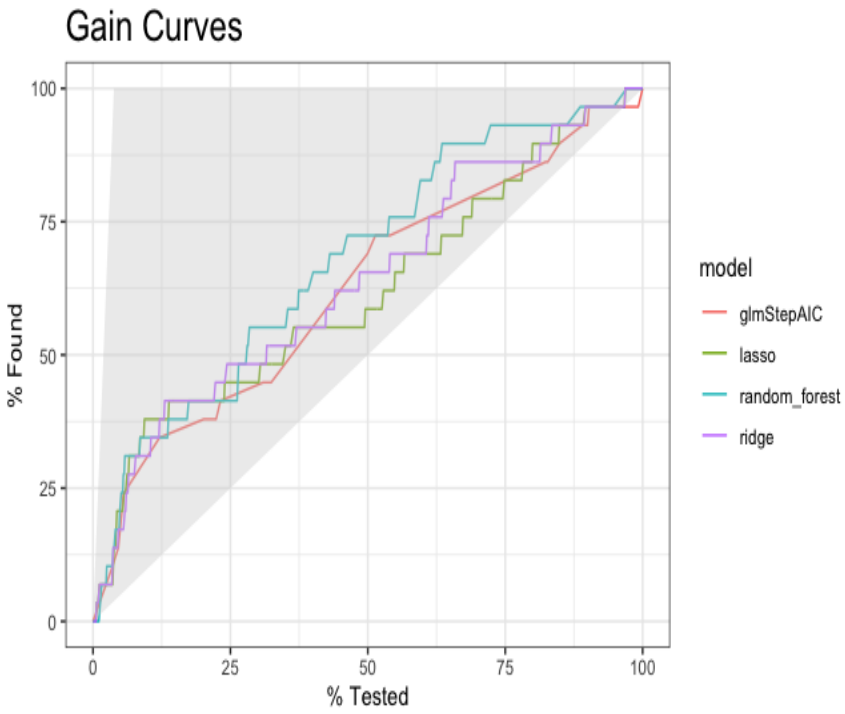


Figure 8. Relapse Gain Curve Model Comparison Optimized for MCC. Copyright 2020 by R.

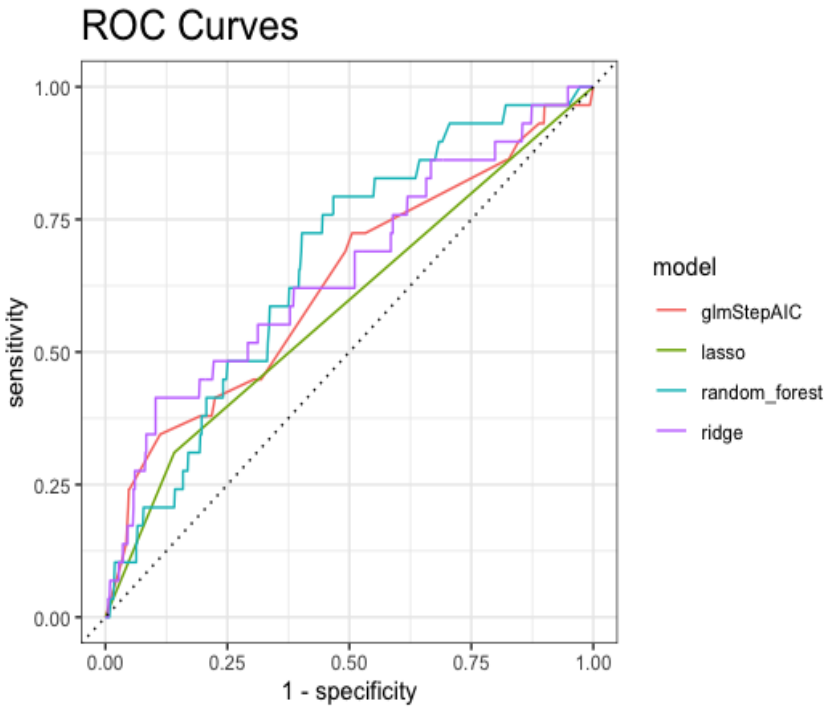


Figure 9. Relapse ROC Curve Model Comparison Optimized for F1. Copyright 2020 by R.

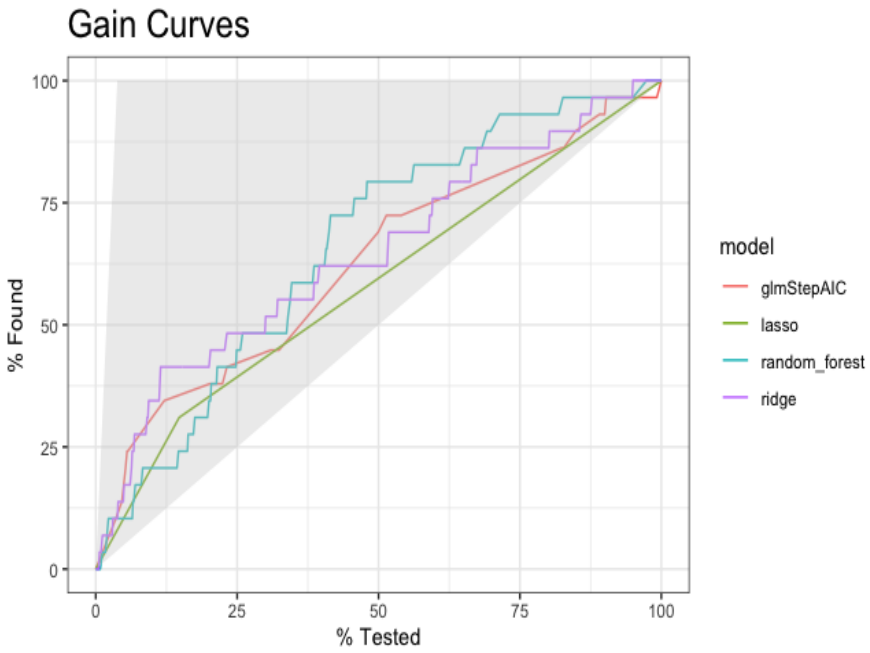


Figure 10. Relapse Gain Curve Model Comparison Optimized for F1. Copyright 2020 by R.

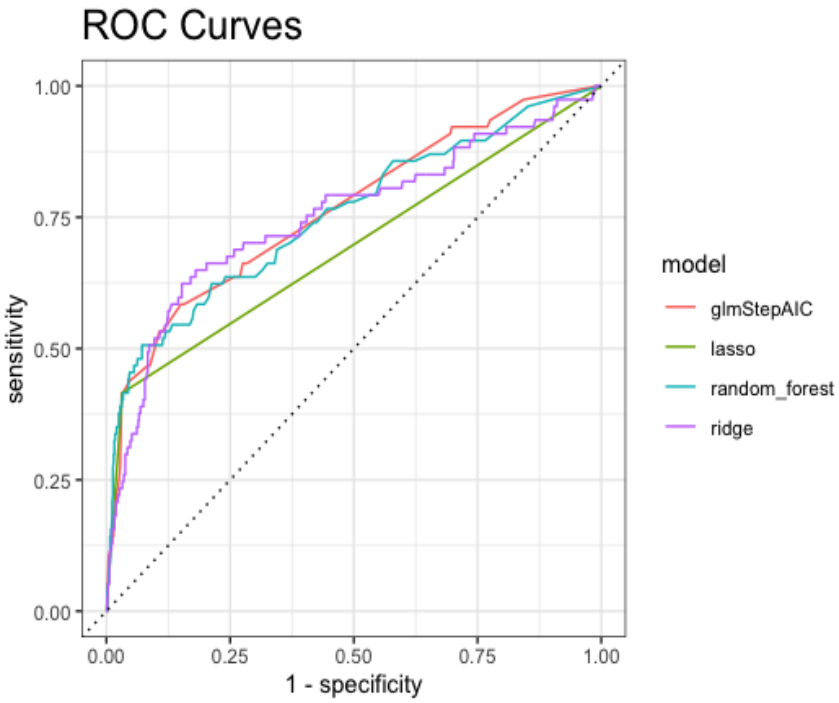


Figure 11. All-Cause Urgent Care ROC Curve Model Comparison Optimized for Accuracy. Copyright 2020 by R.

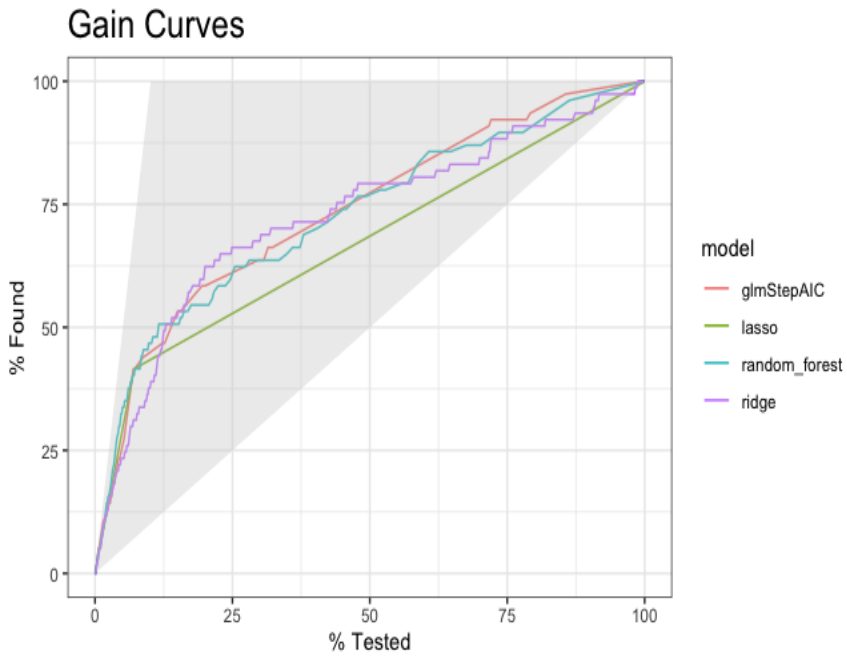


Figure 12. All-Cause Urgent Care Gain Curve Model Comparison Optimized for Accuracy. Copyright 2020 by R.

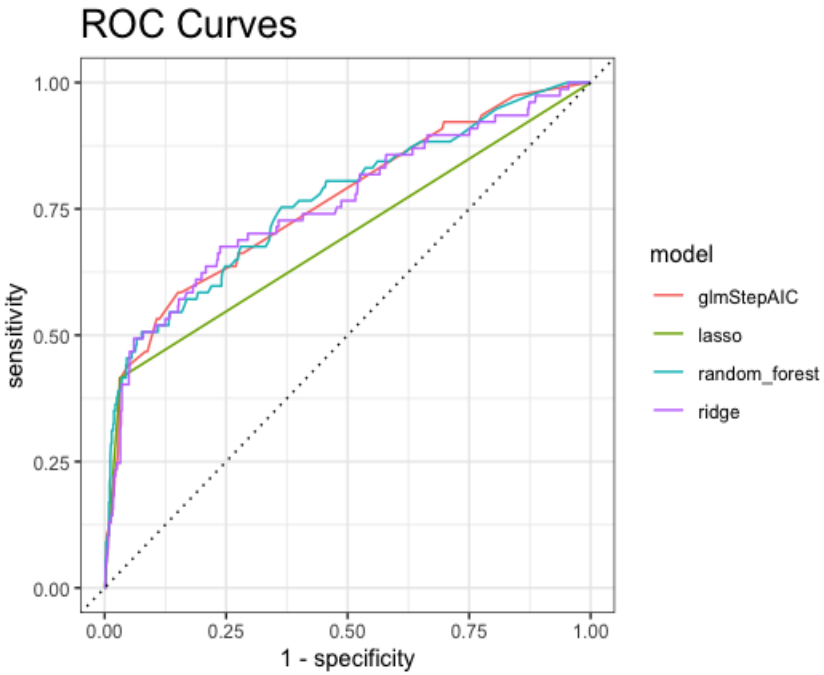


Figure 13. All-Cause Urgent Care ROC Curve Model Comparison Optimized for MCC. Copyright 2020 by R.

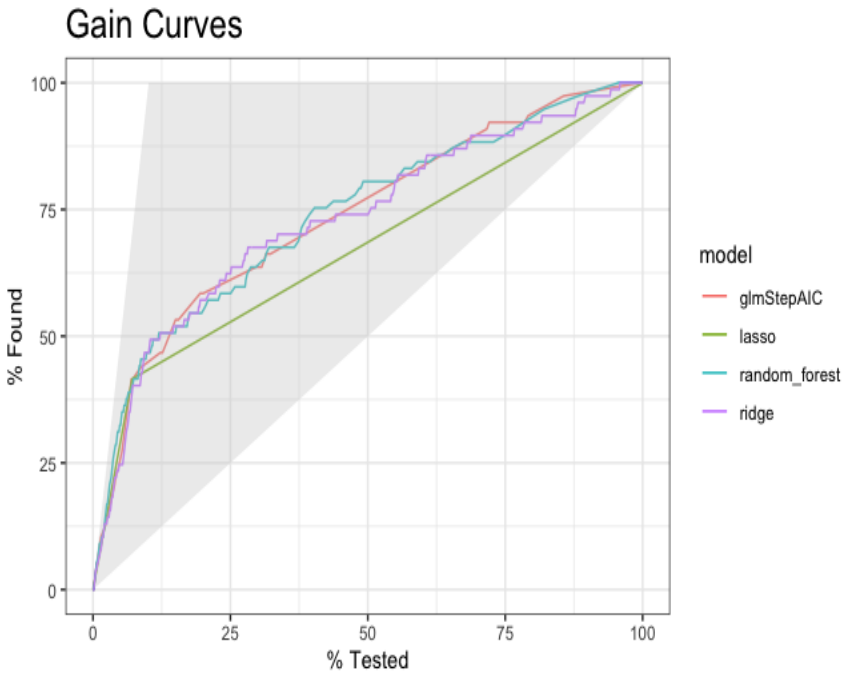


Figure 14. All-Cause Urgent Care ROC Curve Model Comparison optimized for MCC. Copyright 2020 by R.

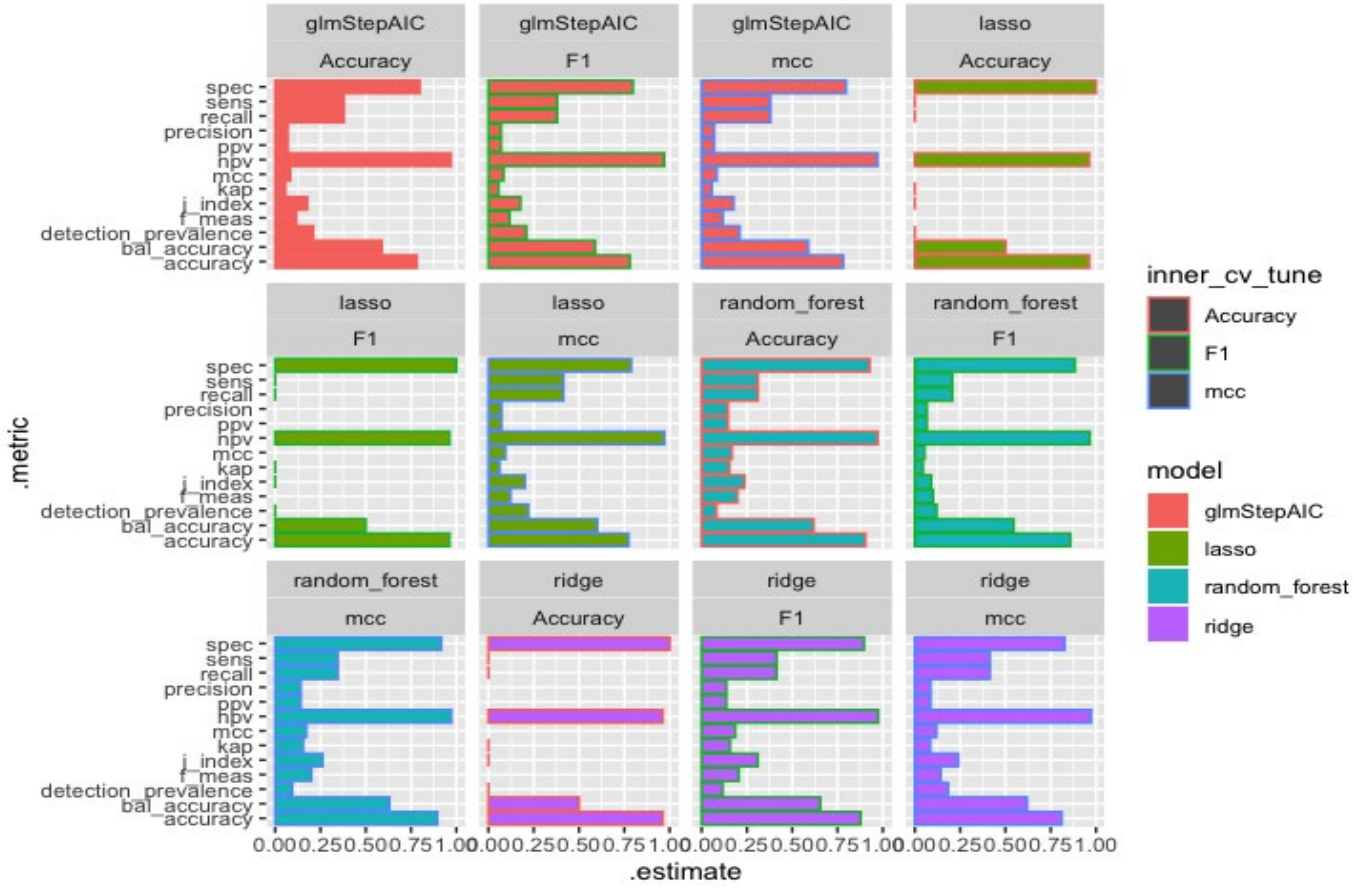


Figure 15. Relapse Model Measures Comparison Including All Optimizations (Accuracy, MCC, F1). Copyright 2020 by R.

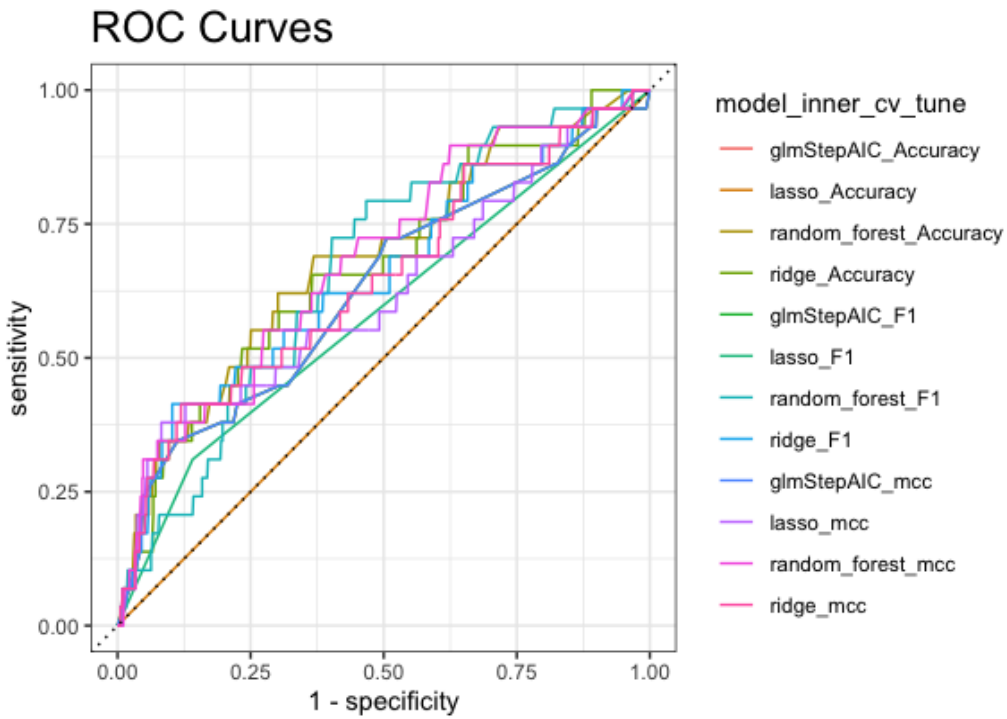


Figure 16. All Relapse Model ROC Curves Including All Optimizations (Accuracy, MCC, F1). Copyright 2020 by R.

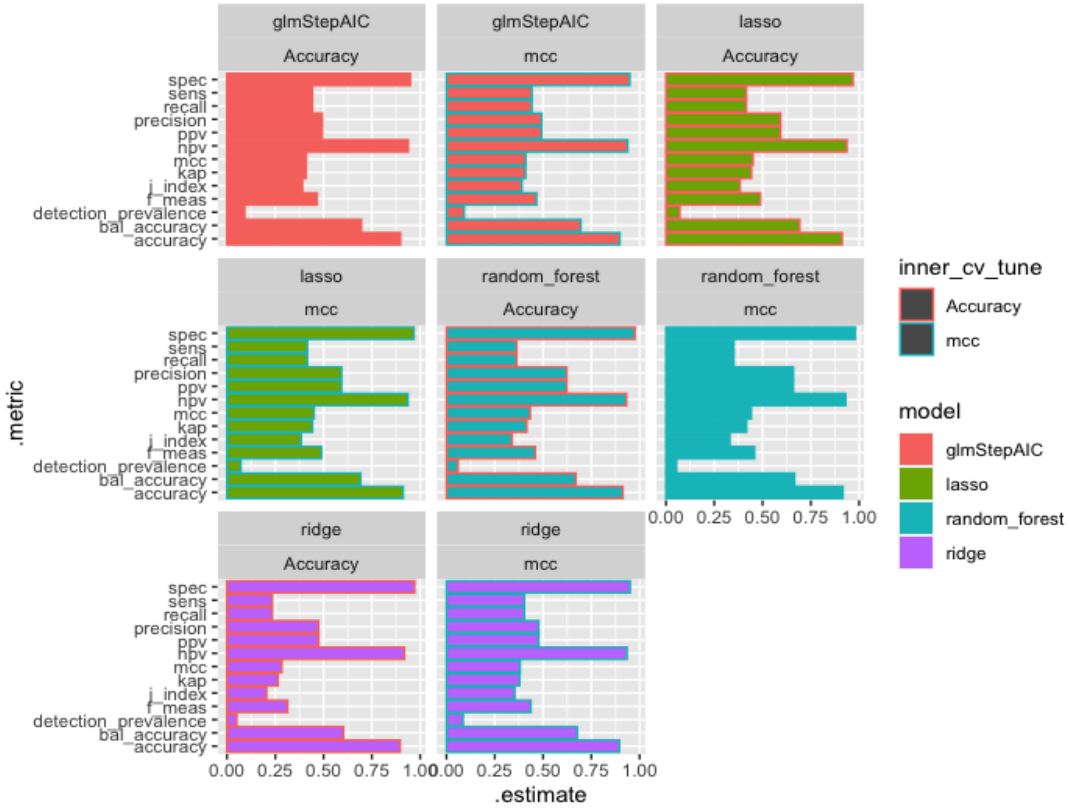


Figure 17. All-Cause Urgent Care Model Measures Comparison Including All Optimizations (Accuracy, MCC). Copyright 2020 by R.

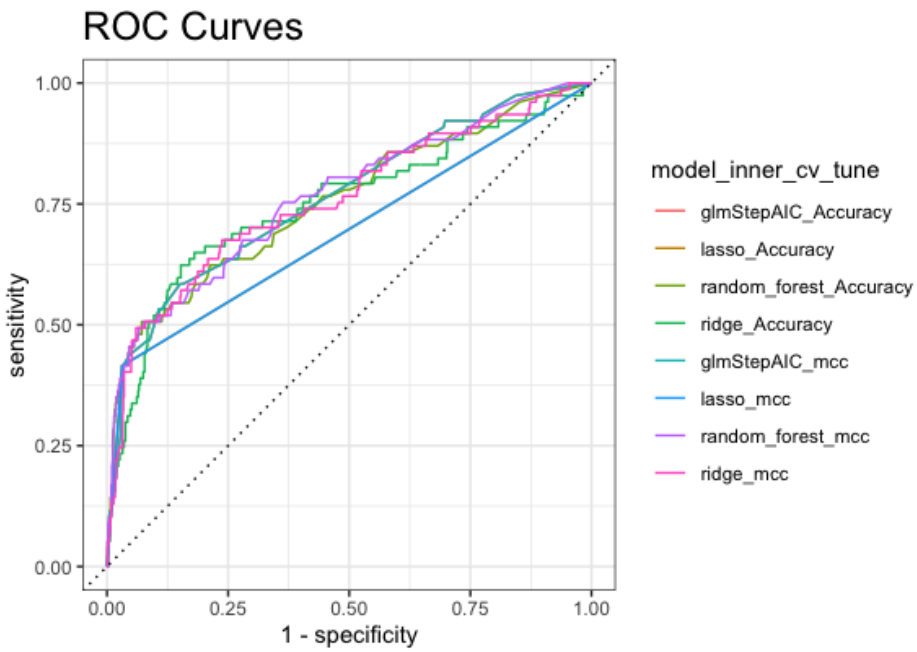


Figure 18. All-Cause Urgent Care Models ROC Curves Including All Optimizations (Accuracy, MCC). Copyright 2020 by R.

Appendix A

IRB Approval

Dartmouth College • Dartmouth-Hitchcock Medical Center
COMMITTEE FOR THE PROTECTION OF HUMAN SUBJECTS
CPHS.Tasks@Dartmouth.edu • 603-646-6482

SOCIAL, BEHAVIORAL, and NON-CLINICAL RESEARCH PLAN
CPHS template v. 10/18/2016

Please complete: CPHS# PI: Brant Oliver (brant.j.oliver@dartmouth.edu)

Important Note: The CPHS Department (Chair & Scientific) Review Form is required with this application. Find the form in the RAPPORT Library or on the CPHS Website.

- **Respond to each item, even if to indicate N/A or not applicable**
- **Attach and/or upload this form as your 'Investigator Protocol' in Rapport**
- **If you are completing this form on a Mac, indicate your answer to any checkboxes by bolding or highlighting, or by deleting any incorrect options.**

1. Introduction and Background

Historically, MS care has been studied and improved at the basic science, individual and population levels of analysis. However, in the new era of healthcare reform these approaches, while necessary and important, will no longer be sufficient. A paradigm shift towards the inclusion of systems-level approaches will be required to study and improve MS care and to demonstrate its value. This new focus derives from three critical developments. First, the *IOM reports*[1] on quality and safety deficiencies in the U.S. healthcare system and the *IHI Triple Aim*[2] have called for a new systems-oriented focus and continuous improvement culture. Second, Wennberg's seminal research on geographic variation,[3] which established the *Dartmouth Atlas of Health Care*,[4] demonstrated that local practice culture and patterns can displace evidence-based care and influence unwarranted utilization and increased costs.[5-6] Finally, the *Affordable Care Act* is driving a shift from productivity to systems-level value-based

reimbursement ("Accountable Care").[7] It follows that a new culture of continuous quality improvement (CQI) will be required to optimize quality and demonstrate the value of MS care at the systems level.

Quality Indicators in Multiple Sclerosis

There has been some discourse regarding quality indicators in MS care. Cheung et al. (2010)[15]offer a substantive matrix of recommended quality metrics developed via a modified Delphi process which are heavily oriented towards clinical outcomes and related process measures (Figures 1 and 2). These span thirteen symptom specific clinical outcome domains (e.g. depression, spasticity, falls, etc.) and nine process specific domains referred to as *"General Health Domains of MS Care"* (e.g. patient education provided at time of diagnosis, provision of community and social resources, etc.). Recent work by the AAN (unpublished draft for public comment)[16] (Figure 3) features a more concise yet similar collection of clinical and process measures focused on the provision of key clinical services and outcomes (e.g. falls screening, falls follow-up, depression screening, etc.), but also includes a functional health (quality of life) category. While encouraging, this work falls short of enabling the global measurement of quality, cost, and value in MS care. For example, cost, utilization, patient experience and satisfaction are not included in any of the current discourse, and disease modifying therapy (DMT) metrics and functional health measures are underrepresented compared to symptom management and clinical care process metrics. Additionally, there is no representation of structural resource measures. These limitations present challenges for cost-effectiveness analyses and the global assessment of the quality and value of MS healthcare services delivery.

Towards Balanced Measurement of Quality and Value in MS

Donabedian's "*Structure, Process, Outcomes*" conceptual framework is well-established and often utilized in healthcare improvement science and can provide a strong foundation for a more comprehensive assessment of system performance in MS healthcare delivery.[17] *Structural measures* include staff, facilities, and resources- the inputs or resources utilized to drive processes and outcomes in system. *Process measures* include practice patterns, e.g. MRI utilization, percentage of patients referred to social services, etc., very similar to many of the process measures listed in Figures 1-2. These measures represent how available structural resources are utilized by the system to generate desired outcomes. Metrics in this domain could span beyond those currently recommended to include a more detailed focus on disease modifying therapy (DMT) treatment initiation and access. *Outcome measures* represent the result of the employment of structural resources and processes to generate desired results. Outcome measures could include many of the metrics in Figures 1-2 such as depression status, annualized relapse rate, etc.

Metrics in this domain could span beyond those currently suggested to include a more detailed focus on DMT treatment related outcomes, such as employment and social security disability status.

Augmenting the Donabedian framework with Nelson's Balanced Measures "*Clinical Value Compass*" framework can further specify *process and outcomes* measurement.[18] Nelson's model has four categories: (1) clinical outcomes (e.g. relapse rate, MRI status, symptoms); (2) functional health (e.g. self-efficacy, quality of life, social security disability status); (3) patient experience and satisfaction; and (4) utilization (e.g. MRI utilization, appointment frequency, ED utilization, etc.). Categories 1-3 are considered to be *quality* measures, and category 4 represents *cost/utilization.* Use of Nelson's framework can enable calculations of *value indices* (Quality/Cost).[18-19]

The proposed study includes the application of this expanded conceptualization of quality, cost, and value measurement utilizing a hybrid combination of the Donabedian and Nelson frameworks. Figure 4 demonstrates how these frameworks can be utilized to fit hypothetical examples of current metrics given in Figures 1-2, as well as additional measures which the current literature does not include.

Driving Systems Level Improvement

Three critical elements are required for improving the quality and value of MS care: (1) systems performance focus; (2) collaboratives; and (3) improvement coaching.

Systems focus refers to a unit of analysis that is aggregated at a higher level than that of the individual clinician (e.g. physician, nurse practitioner, etc.) and which is focused at the level of the service delivery system (e.g. clinic, department, hospital), but at a lower level than that of epidemiological studies of populations (e.g. all MS patients in the United States). This level of analysis is a new focus in MS, which historically has focused on bench science, individual level, or population level analyses. While the recent development of quality metrics in MS is encouraging, its focus on individual clinician (e.g. physician, nurse practitioner) performance is unlikely to generate improvement on its own. The IOM reports (discussed previously) have established that individual efforts to "work harder" or "work better" are usually not enough to generate effective changes without incorporating a system level improvement focus. Additionally, Wennberg's work on local practice culture and geographic variation (also discussed previously) has demonstrated that without the use of benchmarking, transparency and system level improvement efforts, local practice trends remain staunchly resilient, often preventing the timely adoption of evidence based practices and stalling improvement efforts. This suggests that without the inclusion of system-level improvement efforts, the well-intentioned creation of MS quality metrics may only succeed in making hard working clinicians frustrated as they labor in vain to meet expectations that are impossible to reach.

Improvement collaboratives have been shown to facilitate system-level performance improvement[20-21] because they can facilitate benchmarking and transparency of system-level performance, accelerate the rate of learning, and motivate rapid cycle

improvement efforts. Collaboratives are defined as a group of two or more clinical care delivery systems which are focused on the same population and care delivery type (e.g. multiple sclerosis) that agree to share aggregated system level performance data, conduct benchmarking, and engage in learning and improvement together to enable accelerated improvement for all participating sites. The *Cystic Fibrosis Foundation Learning and Leadership Collaborative* (CFF LLC) [11] is an example of a systems-level improvement collaborative. The proposed study includes a small scale improvement collaborative design similar to that of the successful CFF LLC model and follows the general design recommended by the *Institute for Healthcare Improvement (IHI) Breakthrough Series*.[22]

Improvement coaching facilitates, guides, and organizes improvement efforts within healthcare delivery systems. Attempts to support teams in health care improvement have been reported in the literature for several decades and include supportive roles such as coaches, facilitators and helpers.[23-27] Coaching actions include exploring the *context* where the team provides care and services, building *relationships and communication* processes with the improvement team and leaders, offering *helping actions* to support making improvements and reinforcing the improvement process by providing *technical training and support*. According to Gustafson (2013), coaching and improvement collaborative components are equally effective in achieving desired clinical outcomes improvement, and combining them is synergistic and additive.[28] The proposed study recommends the utilization of an experienced improvement coach who has successfully guided improvement teams in a variety of contexts, including the CFF LLC and others, in combination with an improvement collaborative model.

Mesosystem level interaction of microsystems in MS disease modifying therapy (DMT) access

A priority that is shared by persons with MS, clinicians, industry, suppliers, and insurers is timely supply and sustained adherence to DMT (DMT access), which is well established as the basic fundamental backbone of MS care.[40-45] A related critical aspect annual surveillance with brain MRI, which is utilized to monitor MS disease status and detect progression.[46] Current discourse on quality measures in MS addresses DMT and MRI to some degree, but not with specificity.[15-16] It follows that in order to address system level quality of MS care using a CQI collaborative, DMT and MRI utilization should be an initial process performance focus, and that the systems pathway contributing to DMT access be studied. Clinical microsystems concepts and terminology are helpful to employ to articulate this approach. Clinical microsystems represent the smallest possible unit of front-line healthcare service delivery and are the focus point for front-line healthcare improvement work.[9, 19] In MS care, each MS clinical care system (MS clinic or MS center) is a clinical microsystem.

The second order of systems is known as a *mesosystem*, [19] which is defined as two or more interacting microsystems engaged in a shared performance purpose. In the context of DMT access there are three microsystems involved: (1) the MS clinic; (2) the specialty pharmacy; and (3) the insurance provider. Together these three microsystems comprise a "*DMT Access Mesosystem*" (Figure 5). There has not yet been any formal systems level academic or practical study done on this pathway from a healthcare

improvement science perspective. The current state of interaction of these three microsystems appears to be loosely coupled and highly fragmented, with very little shared understanding between microsystem units or focus on collaborative mesosystem level performance. This may result in DMT access barriers at multiple levels (clinic to pharmacy, pharmacy to insurer, etc.). Although the primary focus for the MSCQI Collaborative is on MS clinical microsystems performance improvement, CQI interventions based in those microsystems aimed at improving DMT and MRI performance will need to include the entire mesosystem, including specialty pharmacies and insurers.

Rationale

Systems-based CQI approaches, such as *Lean/Six Sigma*[8] and *Clinical Microsystems*[9] are currently utilized in some healthcare settings across the United States. Regional and national CQI collaboratives utilizing these improvement methodologies have demonstrated significant results. The *Northern New England Cardiovascular Network (NNE) Disease Study Group* has utilized a shared data registry and QI methods to reduce morbidity and mortality across cardiac surgery centers in the northeastern U.S.[10] On an even larger scale, over 110 *Cystic Fibrosis Foundation (CFF)* centers have participated in national level QI collaboratives which utilize a shared systems-level registry and QI methods. The CFF collaboratives have reduced mortality, improved life expectancy, reduced morbidity, and improved a number of process quality indicators.[11-14] Data registries such as *NARCOMS* and the *Slifka Longitudinal MS Study*, have made important strides investigating MS care at the population level of analysis. Unfortunately, these registries are unable to conduct systems-level analyses, and there has not yet been investigation at the systems level regarding geographic variation, quality, and value of MS care. Additionally, there has not yet been an effort of any kind in MS including the use of improvement collaboratives aimed at bettering MS related outcomes, processes, quality or value. This study will establish the first QI collaborative for MS care in the US and will test the effects of QI interventions on selected priority performance indicators.

2. Objectives and Hypotheses

Objective

This is a 3 year study which aims to establish a CQI collaborative of four (4) MS centers/clinics (microsystems), to gather and benchmark systems-level performance across sites and for the collaborative as a whole, and then to study the effect of CQI interventions on improving system-level performance outcomes across microsystems.

Specific Aims

1. To establish systems-level performance indicators by obtaining quarterly performance measures, aggregated by MS center/clinic (microsystem) and the entire *MSCQI Collaborative*.
2. To conduct studies of variation in performance across microsystems and to utilize benchmarking analyses to identify top performers.
3. To study the comparative improvement of selected primary process performance indicators (DMT and MRI process measures) over a 3 year period (Years 2-3) in microsystems receiving IHI Breakthrough Series CQI interventions versus those not receiving CQI intervention.

Primary Endpoint

The primary endpoint for this study is percentage of eligible MS patients on disease modifying therapy (DMT access), which is operationally defined as the total number of eligible patients on DMT/the total number of patients seen per quarter at a participating center for whom DMT is an appropriate treatment option.

Secondary Endpoints

This study will employ a balanced measures conceptual framework (Nelson et al 2004) that is commonly utilized in healthcare quality improvement to measure quality and value at the systems level. From this framework, measures from four functional domains will be employed: (1) clinical process and outcome measures; (2) functional health measures; (3) patient experience measures; and (4) cost and utilization measures.

Secondary objectives are listed below by category:

1) Clinical Outcomes: Depression (PHQ-9), relapse rate, Patient Determined Disease Steps (PDDS).

2) System level (MS center level) aggregate Clinician and Group Consumer Assessment of Healthcare Providers and Systems (CG-CAHPS) scores will be reported quarterly by participating MS centers.

3. Study Design

Describe all study procedures, materials, and methods of data collection:

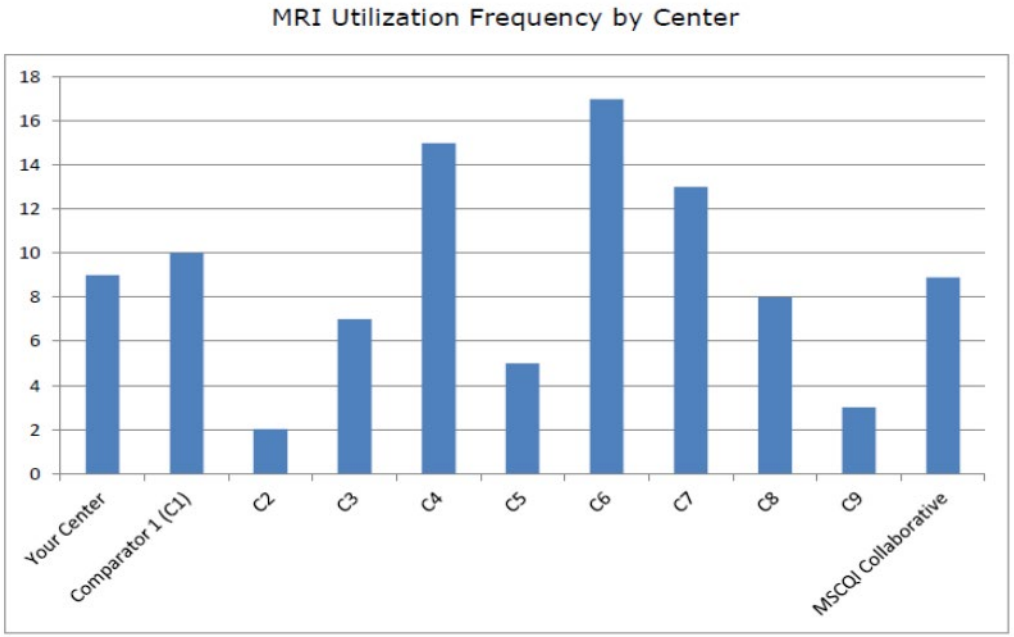
This is a two-part prospective study to be conducted over 3 years with option to extend to 5 years. In the first part (Year 1), we will gather baseline performance data from participating MS clinics (microsystems), create a combined MSCQI systems-level database, and conduct analyses of performance variation and benchmarking (Specific Aims 1 &2). In the second part of the study

(Years 2-3), we will investigate the effect of CQI intervention on primary endpoints and selected secondary measures (Specific Aim 3).

Benchmarking and Longitudinal Performance Monitoring: Years 1-3

Specific aims 1 and 2 will be addressed by quarterly benchmarking of all primary and selected secondary endpoint measures across microsystems throughout the entire duration of the study. Performance data will be aggregated by microsystem and by quarter and submitted electronically via an encrypted and password protected mechanism from each site to the MSCQI Hub Site. No personal identifying information will be collected. Quarterly data collected from sites will be analyzed to create quarterly benchmarking reports for each microsystem and performance reports for the MSCQI collaborative as a whole (basic example of a simple benchmarking data display is given in Figure 8).

Figure 8. Example of a site hypothetical benchmarking report (quarterly MRI utilization)



CQI Step-Wedge/Dynamic Clinical Trial: Years 2-3

The CQI intervention phase of the study will begin in Year 2. We will employ a "step-wedge" design (also known as a Comprehensive Dynamic Trial).[29-30] Compared to a standard RCT or cluster randomized design, the step-wedge design can allow for additional comparisons within microsystems pre- and post- CQI intervention, can allow all participating microsystems to receive CQI interventions, and can better accommodate smaller sample sizes (Figure 9a-c).

Figure 9a. *Step-Wedge/Dynamic Clinical trial (3 Year Trajectory)*

	Benchmarking (Year 1-3)	IHI Breakthrough Series CQI with Improvement Coaching (Year 2)	IHI Breakthrough Series CQI with Improvement Coaching (Year 3)
Center 1	X		
Center 2	X		X
Center 3	X	X	X
Center 4	X		

CQI Intervention

The improvement intervention employed in this study is an *IHI Breakthrough Series* CQI intervention including an improvement collaborative and professional improvement coaching.

IHI Breakthrough Series CQI Intervention with Professional Improvement Coaching

This intervention will be facilitated by clinical teams at participating sites under the guidance of a professional improvement coach utilizing the IHI improvement collaborative model. A hybrid adaptation of the *Cystic Fibrosis Foundation (CFF) Learning and Leadership Collaborative (LLC)* general structure and process as described by Godfrey and Oliver (2014)[11] and the *IHI Breakthrough Series* improvement collaborative model[22] will be utilized. One center will be randomized to this intervention in Year 2.

The selected center(s) for the CQI intervention will form an improvement team consisting of members from the clinic itself. Each team will be advised by the professional improvement coach and instructed in basic improvement methods via a structured curriculum including didactic instruction, supervised application, networking and support provided through on-site meetings, online webinars, and coaching telephone calls (Appendix A). Each improvement team will conduct a system level context assessment of its clinical microsystem[9] with a primary focus on the DMT access and brain MRI monitoring performance and a secondary focus on a selected secondary measure, e.g. patient satisfaction. With guidance from the improvement coach, teams will then formulate small scale interventions and associated real-time monitoring measures aimed at improving process performance for the primary endpoint measures (DMT and MRI). Teams will then pursue rapid cycle implementation, testing, and modification using repetitive *Plan-Do-Study-Act* (PDSA) cycles.[35] Unlike large scale implementation science approaches which follow slower longitudinal trajectories and are more rigidly designed, rapid cycle improvement science approaches such as proposed here are more flexible, adaptive, and short-term. For example, a typical PDSA cycle lasts 2-4 weeks and a typical healthcare improvement trajectory has multiple, successive PDSA cycles, with each cycle adapting and improving upon the last. Successive real-time modifications in the improvement strategy

will be made at the microsystem level by improvement teams based on measured performance feedback. Feedback will be provided via real-time *Statistical Process Control (SPC)* [38-39]measurement (example in Figure 10) conducted directly by improvement teams on the front-line during improvement work and via bi-annual benchmarking reports.

Year 3 Intervention

The "Step-Wedge"/Dynamic Clinical Trial study design will allow for a flexible determination of study exposures in Year 3, which will allow for randomization of a second site to the *IHI Breakthrough Series* intervention(Figure 9)

Year 2 Optional Readiness and Team Assesment Sub Study

The sub study will try to understand the readiness of a clinical improvement team prior to and during improvement activities, including knowledge of quality improvement, systems level variation, readiness to learnand team behaviors (psychology) enabling or disrupting improvement activities. The sub study aims to investigate capability and readiness for engagement in improvement work with the collaborative and to identify potential barriers and facilitators to improvement across the different centers participating in the study. It will complement the MSCQI study focus on system level variation and effectiveness of the QI intervention by providing a descriptive study of the longitudinal readiness and capability of sites to engage in QI, as well as how these characteristics mature over time,

We will employ qualitative interview and quantitative survey instruments designed to assess system and team readiness and capability for improvement work. MS Center team members at sites participating in the MSCQI collaborative will be offered the opportunity to participate in this study during MSCQI study quarterly site visits beginning in Fall 2018 following IRB review and approval. No patient participation or access to electronic medical record information is required. Data will be collected from participating MS center teams at quarterly site visits.

Sub Study Design

Several assessments/tools will be utilized to gain an understanding into the knowledge a team has of Quality Improvement, the team's readiness to learn, and behaviors. The assessments start at all consenting sites during the first quarterly site visit in Year 2 of the MSCQI study and include three time points (beginning of Year 2, beginning of Year 3, and end of Year 3 at the conclusion of the MSCQI study). The assessments will be administered to administrators, staff and providers only. It will not be administered to patients or families and therefore there are no human subjects involved in the study.

The assessments will cover four areas; Knowledge, Readiness to Learn, Behaviors and Team Progress

Knowledge

The Quality Improvement Knowledge Assessment (QIKat) to assess the clinics staff and providers QI knowledge [40]

This tool will be administered to all clinics staff and providers of all of the participating clinics (randomized and control) sites at all three time points. The survey will teak each participant approximately 5 minutes to complete. A Survey Monkey tool will be used to administer the survey (see Appendix E).

Readiness to Learn About QI

Readiness will be assessed for system leaders and improvement teams at all three time points.

Leaders: The clinic leader's readiness to learn is a critical factor in influencing leadership tendency to create conditions for successful improvement for improvement teams. Leader readiness will be assessed utilizing a modified tool "How Groups Learn Continuously" (London, M., 2007). The tool consists of semi-structured questions, with Yes/No and free text options.

This tool will be administered utilizing a Survey Monkey at all three time points. It will be administered to the clinic leaders only as either a direct interview or a Survey Monkey link to the survey. It will take approximately 5 to 7 minutes to complete depending on the comments made by the participants (see Appendix F).

Teams:

The clinics staff's and provider's readiness to learn will also be assessed utilizing a similar tool modified from "How Groups Learn Continuously". This tool also consist of semi-structured questions, Yes/No and free text options. This tool will be administered to all of the clinic staff and providers utilizing a Survey Monkey tool at all three time points. This survey will take the participants approximately 5 minutes to complete (see Appendix G).

Behaviors:

We will evaluate MS Center teams randomized to QI interventions during the MSCQI study, utilizing a structured qualitative assessment tool to describe and characterize behaviors/actions that may have an effect on improvement activities and their outcomes. This is an observational tool and will be administered by the core research team on the teams randomized to the coach supported QI intervention only. To correct for observer bias, at least two observers will utilize the tool to evaluate each team and reconcile by consensus any divergent assessments. The output of the evaluation is to describe categories of behaviors that may enable or disrupt QI activities in the improvement team, and to describe how these behavior characteristics vary across QI teams and change over time. The evaluation will take place during improvement team meetings and/or quarterly site visits during the MSCQI study (see Appendices H, I, and J).

Team Progress

MSCQI sites randomized to QI intervention will undergo brief assessments of quality improvement progress utilizing the IHI "Assessment Scale for Collaboratives" (http://www.ihi.org/resources/Pages/Tools/AssessmentScaleforCollaboratives.aspx). This assessment will be completed monthly by each team that has been randomized to the coach supported QI intervention and by the coach, and the results compared. The tool will take approximately 5 minutes to be complete. A Survey Monkey will be used to administer the assessment. (see Appendix K).

Sub Study Analysis

This study utilizes a descriptive prospective cohort design. Results will be analyzed at the site level with comparison across participating MSCQI sites and compared to the aggregate as a whole.

There will be 3 to 12 possible time pointes depending on the survey type over the length of the research.

Descriptive statistics will be utilized to describe the general performance levels for quantitative measures. We will also perform a correlational analysis between the different measures to assess for interrelationships between these measures. We will use analysis of variance (ANOVA) to assess longitudinal change over multiple time points for these measures to evaluate if statistically significant changes in performance occurred during the course of the study. We will use thematic analysis to develop core themes from the qualitative data collected to organize into potential barriers and facilitators as well as an assessment of overall readiness to engage in improvement work.

Sub Study Progress Monitoring

Participation in the sub-study is at the discretion of each participating center/site and a center's decision to participate in the sub-study will not affect the conduct of the parent (main) study in any way. For the sites participating in the sub study, there will be monthly and quarterly opportunities for the core research team to interact with the participating clinics. At each monthly MSCQI collaborative webinar, the group will review the status surveys completed and outstanding. During the quarterly site visits, the core research team will have the opportunity to review assessments and work to validate data.

Additionally there will be quarterly reviews with the supervising investigator to conduct integrity checks and review the overall study progress and adherence to the approved study protocol.

Study Population

There are two levels of participation in this study: (1) system level administrative; and (2) individual level clinical.

System Level Administrative: The first level of participation is at the system level. It requires the reporting of de-identified administrative system level data aggregated to the system level by the four participating centers. This data will also include covariates such as payer mix and other relevant variables which will be used to adjust between centers. The measures collected for this level of the study will be described subsequently in section 9. This level of the study does not include the collection or use of protected health information (PHI) and will not require informed consent

Individual Level Clinical: The second level of the study is at the individual level. It requires individual patients to complete self-report questionnaires and allow access to their medical records to abstract individual level data. This level of the study will include access to PHI, and will require written informed consent. All data will be de-identified and aggregated to the center level prior to use in the data analyses for this study.

Inclusion Criteria

To be eligible to participate in the *Individual Clinical Level* portion of this study, candidates must meet the following eligibility criteria.

1. Documented diagnosis of multiple sclerosis (MS)
2. Age 18 years or older

3. Ability to understand the purpose and risks of the study and provide signed and dated informed consent and authorization to use protected health information (PHI) in accordance with national and local privacy regulations.

Exclusion Criteria

Candidates will be excluded from study entry if they are unable or unwilling to provide informed consent. Data will not be abstracted from medical records for these patients for the study.

Measures and Data Sources

This study will employ a comprehensive system level performance measurement strategy featuring quarterly reporting, semi-annual benchmarking data feedback to sites, a one year baseline assessment period, and a two year evaluation of QI intervention effects.

Three data sources will be utilized for this study: (1) center level administrative data; (2) individual level data abstracted from medical records and subsequently de-identified and aggregated to the center level; and (3) individual self-report data connected by electronic data capture (EDC) methods that will subsequently be de-identified and aggregated to the center level. Variables for the study are given in Table 2 below and are listed by Nelson's domains (clinical outcomes, functional health outcomes, patient experience, and utilization/cost). Table 2 lists data sources, the priority of the data collection method (core versus exploratory) and the type of endpoint (primary versus secondary). Clinical and functional health measures, which are derived from medical record abstraction (MRA) or self-report questionnaires, require informed consent whereas patient experience, utilization, and covariate measures, which are derived from administrative data, will not require consent (see Table 2a). *Table 2. MSCQI Collaborative Measures*

Category	Measure(s)	Type	Source	Frequency	IC

Population Characteristics	Race/Ethnicity, Socioeconomic status Primary language Marital status Education Employment Co-morbidities, Concomitant meds, Current/previous DMTs	Supplemental	PRO Portal	Annually	Yes
Clinical	DMT Access (% on DMT)	Core (Primary)	Medical record abstraction (MRA)	Quarterly	Yes
Clinical	% MRI in past year	Core	MRA	Quarterly	Yes
Clinical	Relapses	Core	MRA & PRO Portal	Quarterly	Yes
Clinical	PHQ-9	Core	PRO Portal	Quarterly	Yes
Clinical	Vitamin D Level	Core	PRO Portal	Annually	Yes
Clinical	PDDS	Core	PRO Portal	Semi-Annually	Yes

Functional Health	Neuro-QoL Ability to Participate in Social Roles and Activities Lower Extremity Function Upper Extremity Function Stigma Satisfaction with Social Roles and Activities Sleep Disturbance Communication	Core	PRO Portal	Semi-Annually	Yes
Functional Health	Neuro-QoL Anxiety Cognitive Function	Core	PRO Portal	Quarterly	Yes
Functional Health	WPAI	Core	PRO Portal	Semi-Annually	Yes
Functional Health	PROMIS Fatigue SF	Core	PRO Portal	Quarterly	Yes
Patient Experience	BAI	Core	PRO Portal	Quarterly	Yes
Patient Experience	CG-CAHPS My Health Confidence	Core	Administrative	Annually	No

	Your Opinion Matters				
Patient Experience Adherence	TSQM-9	Core	PRO Portal	Semi-Annually	Yes
Utilization	MRI	Core	Administrative	Quarterly	No
Utilization	Office Visits ED Visits Hospitalization	Core	Administrative	Quarterly	No

Table 2a. Descriptive & Covariate Measures: Collected Annually by Center (Administrative)

Category	Measure	Information
Population Core Demographics	% Female	Gender
Population Core Demographics	% RRMS	MS Disease Type
Population Core Demographics	Age distribution	Age
Region	State	Geography
Payor Mix	% Private Insurance	Insurance coverage
Payor Mix	% Government Insurance	Insurance coverage

Providers (MD, DO, NP, PA)	#FTE	Capacity
Nursing (RN, LPN)	#FTE	Capacity
Social Worker/Mental Health	#FTE	Capacity
Rehab Providers (PT, OT, etc.)	#FTE	Capacity/Access
Provider Visits Available	Average #visits available per month	Capacity/Access (Potential)
Certification	NMSS or CMSC certified	Recognition as Center of Excellence for MS care
Clinical Setting	Academic, Community Hospital, Private	Practice type
Clinical Setting	Urban, Rural	Community setting

Data Collection and Processing Pathway

Data to be collected in this study will be abstracted quarterly in de-identified form via data downloads from electronic medical records and administrative records from participating centers and from the PRO Portal as previously described (see Figures 6-7). EMR Data will be entered by each site into a secure Redcap database maintained on the Dartmouth Hitchcock server. No PHI will be entered into this database. All PHI will remain at the specific sites. Once the data has been aggregated into the Redcap database it will be submitted in a highly encrypted, aggregate form to a MSCQI hub site database along with the PRO data for management and analysis. *Jefferson School of Population Health (JCPH)* will provide data management and analyses for this study. A hub site database will be constructed and managed by *JCPH* using its own resources and secure firewalls. Security standards for JCPH will be

assured at computing levels required by Dartmouth Hitchcock Medical Center. Site sub-investigators and/or their designees, and the MSCQI research coordinator will facilitate data abstraction and secure transmission to the hub site database with assistance and oversight provided by the study methodologist and the principal investigator.

Participating Sites and Key Personnel

This study will be facilitated by four entities: (1) the core research team (*MSCQI Hub Site*); (2) an independent database management site (*JCPH*); (3) MSCQI sites (microsystems) and sub-investigators; and (4) the Research and Improvement Advisory Committee (RIAC).

(1) MSCQI Hub Site and Core Research Team Personnel

Leadership and management for the study will be centralized at the Dartmouth Population Health Collaboratory research hub site as previously described in Section 1. The core research team will consist of the principal investigator (PI), the methodologist, the professional improvement coach, and the research program manager/study coordinator.

- *Principle Investigator (Brant Oliver, PhD, MS, MPH, APRN-BC, MSCN):* See Section 1.
- *Co-Investigator/Improvement Coach (Randy Messier MT, MSA, PCSH CCE):* Mr. Messier will serve as improvement coach and co-investigator for this study. He is an experienced professional improvement coach with management and administrative training and is also certified in Patient-Centered Specialty Home (PCSH CCE). He has served successfully as a professional improvement coach for many hospitals and improvement collaboratives in the United States and Canada, including the CFF LLC national collaborative (during which time he has worked closely with Dr. Oliver)[11] and multiple state-level improvement collaborations in Vermont, including recent work in the *Optimizing Laboratory Testing Collaborative* .

- *Research Program Manager and Study Coordinator (Amy E. Hall, MS)*: Ms. Hall will serve as research program manager and study coordinator. She has extensive experience in program management, study coordination and data abstraction. Her duties will include coordination between the hub site and participating MSCQI sites, regulatory and budgetary coordination, logistics, and data abstraction (including travel to MS centers for data abstraction).

(2) Independent Data Management and Data Analytics Center (*Jefferson College of Population Health Innovation*)

The MSCQI Collaborative study will maintain its research database at a site independent of the funding source and the study PI. This is of particular importance in industry-funded and investigator-initiated studies to reduce actual and/or perceived risks of bias in design and conduct of the study and analysis of the data. For this reason it is proposed that Jefferson College of Population Health Innovation (JCPH) provide data management for this study and that this study and that this aspect of the study be overseen by Alexis Skoufalos, EdD. The MSCQI hub database will be constructed and managed by JCPH.

(3) MSCQI Sites (Microsystems) and Sub-Investigators

Four (4) multiple sclerosis centers/clinics (microsystems) from the eastern United States will participate in the MSCQI collaborative study. Oversight for each participating site will be provided by a study sub-investigator who will serve as a site principal investigator. Site investigators are experienced MS clinicians and/or researchers who have worked with the study PI (Dr. Oliver) in the past and have the capability to manage and coordinate activities at their respective sites to accommodate the needs of the proposed study, including data abstraction and reporting, and engagement in CQI efforts during the intervention phase of the study. Total panel size for all four sites combined is approximately 5,600 MS patients (N=5,600). Participating sites represent urban and rural settings, academic and private practice contexts.

- *MGH Multiple Sclerosis Clinic (Site Investigator: Eric Klawitter, MD, M.Sc.):* The MGH MS Clinic is an urban academic MS center in Boston, MA that is affiliated with the Partners Healthcare system. It follows approximately 1,000 MS patients and is heavily involved with MS clinical trials. Dr. Klawitter is a NMSS fellowship trained MS neurologist and research scientist with expertise in MRI imaging. He directs the MGH MS Clinic.

- *University of Vermont (UVM) Multiple Sclerosis Center (University of Vermont Medical Center (Site Investigator: Andrew Solomon, MD):* The UVM Multiple Sclerosis Center is a rural academic MS center in Burlington, VT which follows approximately 1,500 MS patients. It is affiliated with the University of Vermont (UVM) Medical School, and is heavily involved with clinical trials research.

- *Neurology Associates Multiple Sclerosis Center of Greater Orlando (Site Investigator: Patricia Pagnotta, MS, APRN-BC, MSCN):* The multiple sclerosis practice at Neurology Associates of Greater Orlando represents a large private MS clinic, following approximately 1,000 MS patients. The MS practice there has an extensive MS clinical trials research program paralleling similar programs in academic MS centers. Patricia Pagnotta is an experienced certified MS nurse practitioner in the practice and is closely involved with the MS research program.
- Concord Hospital Multiple Sclerosis Specialty Care Clinic (Site Investigators: Ann Caobt, DO and Jennifer Taylor, ARNP): The multiple sclerosis practice at Concord Hospital represents a small private hospital in a rural community. The clinic follows approximately 1200 MS patients.

4. Analysis

Describe any qualitative tests and measures as well as quantitative methods:

The data analytic plan for this study is intended to reflect, to the degree possible, the actual effects of system performance on population health. The analysis population will include all administrative data collected at the system level combined with all individual level data collected from all consenting individuals, aggregated up to the systems level. Statistical tests will be conducted using *IBM SPSS Version 22* (Chicago, IL) and/or *STATA* (Statacorp Inc., College Station, TX) software with alpha set at $p < .05$.

Prior to analysis, all responses will be examined for accuracy of data entry, missing values, and fit between their distributions and the assumptions of the statistical analysis including normality, homogeneity of variance, linearity, and colinearity. Violations of the aforementioned assumptions will lead to transformations (square root, logarithm, or inverse) to reduce skewness, reduce the number of outliers, and improve the normality, linearity, and homoscedasticity of residuals. Subsequently, specific descriptive and analytic examinations will be conducted by specific aim as described below.

Specific Aim 1: To establish systems-level quality and value (i.e. cost-effectiveness, efficiency) performance levels by obtaining quarterly performance measures, aggregated by MS clinic (system) and the MSCQI Collaborative as a whole.

Appropriate descriptive statistics will be conducted to describe basic system-level variation in primary endpoint measures (DMT and MRI) and selected secondary endpoint measures from each of the major balanced measurement domains (clinical, functional health, patient experience, and utilization). Measures will be adjusted based on appropriate system level structural and individual level demographic characteristics which demonstrate statistically significant relationships with the dependent variable in univariate analyses.

Specific Aim 2: To conduct studies of geographic variation in performance across MS clinics (microsystems) and utilize benchmarking analyses to identify top performers.

A hierarchical linear model (HLM) for performance levels is recommended for Aim 2. The HLM will incorporate the performance measures and the effects of geographic location. To assess the importance of geographic variables, three types of models will be compared: a model with individual level variables only; a model with system level geographic effects that do not interact with person attributes; and a full model, allowing for geographic level random effects that differ by site. Both of the primary endpoint measures and at least one of each category of quality measures will be included in benchmarking analyses: (1) clinical measures (e.g. EDSS); (2) functional health measures (e.g. Neuro-QOL); and (3) a patient satisfaction measure (CG-CAHPS). The *Available*

Benchmarks of Care (ABC) Benchmarking Method[37] may also be utilized to benchmark variables that demonstrate substantive variation in key process or outcomes performance and to identify top performing sites by performance category.

Specific Aim 3: To test the effect of rapid cycle CQI interventions in improving selected quality and value outcomes in participating sites

Standard monitoring of all measures will continue throughout the intervention phase in Years 2-3 similar to that of Year 1. Longitudinal quantitative analyses, site specific real time assessments using statistical process control (SPC), and qualitative analyses can be conducted to determine improvement effect in intervention sites compared to controls and compared to each other. See sections 11.4-11.6 below for detailed descriptions of analyses for this specific aim

Primary Endpoint Analysis

This section is relevant to the analysis of Specific Aim 3 as previously discussed above. Longitudinal time series regression analyses will provide a robust assessment of overall intervention effects on DMT treatment access. DMT access will be treated as the DV and compared between baseline and quarterly intervention time periods during Years 2-3, overall and stratified by intervention and benchmark. We will use Chi-square tests for categorical data and Student t-tests for continuous data. We will use multilevel XTME Poisson regression clustering to the clinic/center level to calculate adjusted risk ratios (RR) with 95% confidence intervals (95%CI) of the selected DV measures between the intervention and baseline periods adjusting for important system level structural and individual level demographic characteristics. To demonstrate temporal trends in performance, we recommend interrupted time series analyses of quarterly performance rates adjusting for covariates in the Poisson model.

Secondary Endpoints

Secondary endpoints, as described previously, will be analyzed using similar methods to the primary outcome variable. Longitudinal time series regression analyses will provide a robust assessment of overall intervention effects on secondary outcome

variables. These variables will be treated as DVs and compared between baseline and quarterly intervention time periods during Years 2-3, overall and stratified by intervention and benchmark. We will use Chi-square tests for categorical data and Student t-tests for continuous data. We will use multilevel XTME Poisson regression clustering to the clinic/center level to calculate adjusted risk ratios (RR) with 95% confidence intervals (95%CI) of the selected DV measures between the intervention and baseline periods adjusting for important system level structural and individual level demographic characteristics. To demonstrate temporal trends in performance, we recommend interrupted time series analyses of quarterly performance rates adjusting for covariates in the Poisson model.

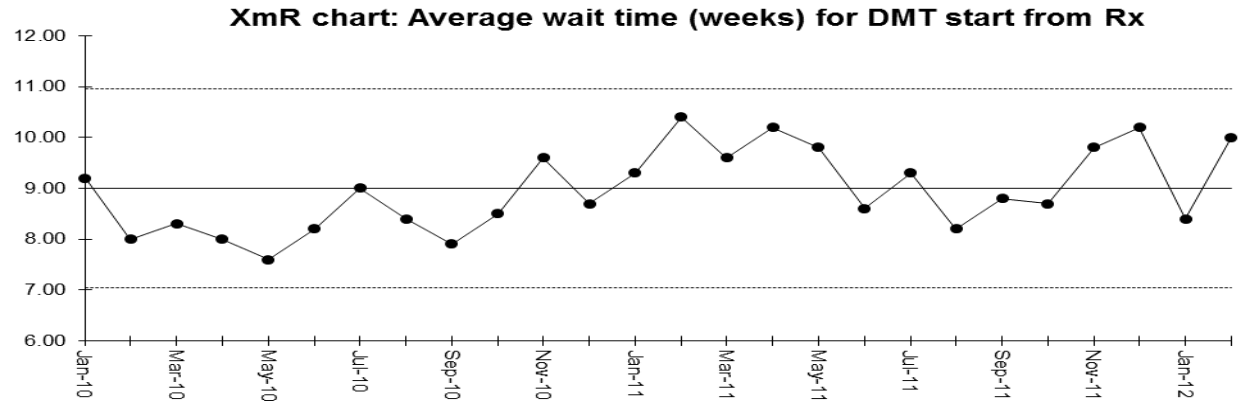
Interim Analyses

Two types of interim analyses will be conducted during the study: (1) benchmarking analyses (see Specific Aim #2 above); and (2) Statistical Process Control (SPC) analyses (see Specific Aim #3 above). These sections are repeated below.

1) *Benchmarking Analyses (Specific Aim #2):* Benchmarking analyses will be conducted semi-annually and results reported as feedback to sites to contribute to the overall function and improvement mission of the MSCQI Collaborative. A hierarchical linear model (HLM) for performance levels is recommended for Aim 2. The HLM will incorporate the performance measures and the effects of geographic location. To assess the importance of geographic variables, three types of models will be compared: a model with individual level variables only; a model with system level geographic effects that do not interact with person attributes; and a full model, allowing for geographic level random effects that differ by site. Both of the primary endpoint measures and at least one of each category of quality measures will be included in benchmarking analyses: (1) clinical measures (e.g. EDSS); (2) functional health measures (e.g. Neuro-QOL); and (3) a patient satisfaction measure (CG-CAHPS). The *Available Benchmarks of Care (ABC) Benchmarking Method*[37] can be utilized to benchmark variables that demonstrate substantive variation in key process or outcomes performance and to identify top performing sites by performance category.

2) *Statistical Process Control (Specific Aim #3):* SPC measurement methods[38-39] will be utilized by centers randomized to the IHI Breakthrough Series QI intervention. SPC analyses will be developed by front-line improvement teams during rapid cycle change interventions (PDSA cycles) to assess short-term, context- specific performance metrics and to adjust real-time CQI intervention efforts accordingly in order to maximize longitudinal improvement performance. An example of a SPC chart is given in Figure 10.

Figure 10. A Statistical Process Control (SPC) Chart used in real time CQI work by improvement teams (for a site level process measure related to DMT access primary outcome)



5. Study Progress Monitoring

Note: appropriate monitoring may include periodic assessment of the following:

- data quality
- timelines
- recruitment and enrollment

Provide a description of the methods which will be used to determine the progress of the study, including periodic assessments of data quality, timelines, recruitment, and enrollment as appropriate:

Study monitoring will be conducted on two levels: (1) investigator level; and (2) via a Research and Improvement Advisory Committee (RIAC).

Investigator Level Monitoring

The Dartmouth research hub site will monitor overall study progress in real time with adherence to quarterly time points established by the study schema given below. This schema will provide the study timeline, which provides quarterly time points for data collection and transmission to the data analytics center (TBD), at which time a quarterly assessment of recruitment and enrollment will also be ascertained. The study investigator will meet regularly with the data analytics center at a minimum of once monthly for ongoing work on data collection and analysis, and this work will also include a quarterly assessment of data quality.

Table 1. Study Activities Schema

	Q0	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**	**Q11**	**Q12**
Start of Quarter	**-**	**7/1/17**	**10/1**	**1/1/18**	**4/1**	**7/1**	**10/1**	**1/1/18**	**4/1**	**7/1**	**10/1**	**1/1/19**	**4/1**
Deadline for Deliverables	**12/31**	**10/1**	**1/1**	**4/1**	**7/1**	**10/1**	**1/1**	**4/1**	**7/1**	**10/1**	**1/1/19**	**7/1**	**6/30**
Start-up, Contracting	**X**												
Internal Review Board (IRB) initial submission and reviews	**X**				**X**				**X**				**X**
Annual Report to Funder					**X**				**X**				**X**

Measures collected (All Sites)		**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**
Benchmarking Reports to Sites			**X**		**X**		**X**		**X**		**X**		**X**
CQI Intervention (Step-Wedge Design)					**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**
Coaching Site Visits				**X**	**X**			**X**	**X**			**X**	**X**
Coaching Calls with Teams					**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**
Improvement Skills Webinars					**X**	**X**	**X**	**X**		**X**	**X**	**X**	**X**
On-Site Sessions					**X**				**X**				
RIAC meetings			**X**		**X**		**X**		**X**		**X**		**X**

Research and Improvement Advisory Committee (RIAC)

The *RIAC* will provide balanced advisement, transparency and monitoring of study activities by key stakeholders and technical experts. The RIAC will be chaired by a MS clinician or researcher selected by the Principal Investigator (Dr. Oliver), and will include at least one of each of the following additional members (also selected by the PI): (1) at least one person who has MS; (2) a scientific representative designated by the funder; (3) a healthcare quality improvement specialist and/or researcher; (4) a representative from a DMT specialty pharmacy; (6) a second MS specialist clinician or researcher; and (5) a representative from an insurer/payer (if available).

The RIAC will meet biannually with the Principal Investigator and other core research personnel as needed to discuss study progress and recommendations for the ongoing conduct of the study. RIAC members will be bound by a confidentiality agreement and must complete online *CITI training for human subjects research*. The RIAC will be of particular importance in assuring an appropriate practical balance between stakeholder needs, ethical standards, and risk of stakeholder bias in industry-funded and investigator-initiated research focused on the assessment and improvement of systems-level performance in MS centers/clinics. Maintaining this balance will be critical to the investigators and the funder in order to cultivate a high standard of credibility concerning the conduct of the study and the integrity of its methods and results.

6. Risks & Benefits

Note: Risks may be physical, psychological, social, legal, economic, to reputation, or others.

a. Describe any potential risks, their likelihood and seriousness:

Because this study does not employ any invasive procedures or treatments, the risk to participants for participating in the study is assumed to be minimal and no greater than that engendered by participating in routine MS care. This study will include the collection of personal identifying information (PHI) which introduces associated risks, but these are estimated to be low as will be discussed in later sections addressing data security. Study procedures will not alter standard multiple sclerosis care for participants or influence the patient-provider interaction or relationship in any way. Data collected for this study will be managed by experienced personnel at a designated independent data analytics center (JCPH).

Designated Data Analytics Center (JCPH)

Data to be collected in this study will be abstracted quarterly in de-identified form via data downloads or manual abstraction from electronic medical records and administrative records from participating centers and from the PRO Portal as described below and in data flow process Figures 6-7. Data will be submitted in aggregate form from the Dartmouth hub site to a MSCQI database for management and analysis. *Jefferson College for Population Health Innovation* will provide data management and analyses for this study. A hub site database will be constructed and managed by JCPH using its own resources and secure firewalls. Site sub-investigators and/or their designees, and the MSQCI research coordinator will facilitate data abstraction and secure transmission to the hub site database with assistance and oversight provided by the study methodologist and the principal investigator.

JCPH was purposively recruited to provide research methodology and data analytics services for the MSCQI Collaborative study for the following reasons: (1) significant past history of working with academic institutions and multiple industry entities without a predominant interest or affiliation with any particular entity; (2) significant expertise in the organization and its leadership in population health research (including patient reported outcomes, quality of life research, and healthcare quality research); (3) capacity to serve as an independent data repository and data analytics entity functioning separate from the MSCQI Collaborative entities and the funding source, thus optimizing the integrity of data management and reducing threats to objectivity posed by stakeholder (funder, study site, etc.) interests.

Patient Reported Outcomes (PRO) Portal Description

The PRO Portal system (see Appendices C and D) will include a secure browser and device compatible system with the ability to register role based users, ability to deploy validated electronic surveys, develop surveys, a graphical view module containing descriptive statistics for real-time acquired data, a system dashboard containing user characteristics, and a timeline module to document key historical events for groups of participants. Participants will use system to enter health data, complete validated electronic surveys, document event driven events, i.e. MS relapse, concomitant medications, co-morbidities, etc. Investigators will login to review de-identified aggregated data in a graphical view module for aggregate de-identified data downloads for study-related data analyses. All PHI stored in the PRO Portal is 128 bit encrypted and fully compliant with current HIPAA standards (see Appendix

C). Access to encrypted PHI in the PRO Portal will be upon secure log-in with the operations system by authorized users only. Participants will not be identified by name in de-identified data used for study analyses or publications and data collected will be used for research purposes only.

b. Confirm that risks to subjects have been minimized, by use of procedures which are consistent with sound research design and which do not unnecessarily expose subjects to risk:

As discussed in Part A above, actual risks to participants are assumed to be minimal and no greater than routine multiple sclerosis care and because de-identified data will be collected, confidentiality risks due to study participation are minimized. Additionally, the use of a step-wedge research design allows for rigorous empirical investigation of the research question utilizing a robust set of measures but with a low time duration and limited number of participating MS Centers. Additionally, the study does not influence the patient-provider relationship or selection of multiple sclerosis medical treatments in any way. This combination of study design characteristics effectively minimizes participant burden and risk.

c. Describe why all the risks to subjects are reasonable in relation to both anticipated benefits and the knowledge expected to be gained from the study:

As described in Part A and B above, risks associated with study participation are anticipated to be minimal and no greater than that associated with standard multiple sclerosis care. The anticipated benefit of this study is the establishment of the first national quality improvement collaborative for multiple sclerosis in the United States and initial research on system level performance variation in multiple sclerosis care and the comparative evaluation of quality improvement interventions aimed at bettering multiple sclerosis care outcomes and care quality. This study could greatly benefit the multiple sclerosis care community by generating new knowledge about best practices and new approaches to improving system level performance in managing this very common and costly

chronic illness population. Finally, this study may also establish key initial findings demonstrating the feasibility of an improvement research collaborative for multiple sclerosis and a framework for subsequent research and development in this area.

7. Unexpected Events or Incidental Findings

Note: It may be important to consider the potential for certain unanticipated events to occur, for example:

- finding an anomaly in a MRI
- discovering child abuse
- causing distress in interviews of a sensitive nature

Describe potential events and provide a plan of action:

Because this study does not employ any invasive procedures or treatments, the risk to participants for participating in the study is assumed to be minimal and no greater than that engendered by participating in routine MS care. However, throughout the course of the study, every effort must be made to remain alert to possible adverse events (AEs). If an AE occurs, the first concern should be for the safety of the subject. If necessary, appropriate medical intervention should be provided. All participants in the study will be given a study information sheet which will contain contact information for the study investigator. Any untoward or unfavorable medical occurrence in participants, including any abnormal sign, symptom, or disease, temporally associated with the subject's participation in the research, whether or not considered related to the subject's participation in the research, will be termed an "adverse event" and reported.

Adverse events will be evaluated by the investigator or his medically qualified delegate to determine if the adverse event represents an unanticipated problem. Adverse events that meet the following criteria will be deemed unanticipated problems and reported to the IRB:

- The adverse event is unexpected

- The adverse event is related or possibly related to participation in the research
- The adverse event suggests that the research places subjects or others at a greater risk of harm than was previously known or recognized

8. Deception

Does any part of this study involve deception or withholding of information from participants?

☐ **Yes** ☒ **No**

If Yes, provide an explanation which addresses the following:

- A description of the deception being used
- Why the deception is necessary
- A plan for debriefing, or providing subjects with the pertinent information after participation

9. Equitable Participant Selection

a. Estimated number of participants at Dartmouth CPHS reviewed sites:

A sample size of 2,000 participants (cumulatively from all four participating sites) is estimated based upon the proposed study measures and data analytic plan. Two of the participating sites (Orlando and MGH) have ceded IRB review to Dartmouth (representing approximately one half of the 2,000 total, or 1,000). The other two participating centers (University of Vermont and Concord Hospital, representing the other 1,000 participants, have elected to undergo separate IRB review by their respective institutions).

b. Provide a justification of the proposed sample size

This study aims to investigate and improve "real-life" MS care delivery at a variety of MS centers (microsystems) across the United States. A convenience sample of four (4) MS centers/clinics (sites) that have expressed interest in participating in this study have been recruited, following approximately 5,600-6,000 persons with MS. Actual sample size (total number of participants required) for data analyses for main effects based on our proposed data analytic plan is determined by the final number of variables selected for inclusion in the final regression model for main effects analysis. This will be determined based on univariate analyses of significant factor contributors to outcomes for each of the component measures as well as factor analyses which will identify significant clusters of variables which can be represented by the most significant contributors in each cluster (thus minimizing sample size demand). Given that baseline data will be required to conduct univariate analyses and factor analyses to calculate the actual sample size for final main effects regression models, we used standard conventions for multiple regression modeling[36] to estimate a maximum sample size for conducting primary analyses with acceptable power for final main effects regression models containing ten variables and repeated measures in a step-wedge research design. We will use this as our final modeling constraint, i.e. a maximum of ten "highest priority variables" for each main effects primary endpoint regression analysis. We estimate that a sample size of approximately 2,000 participants (500 per recruitment site) will be required for acceptable power for main effects analyses given these constraints. To meet the recruitment goal of 2,000 participants, we will need to maintain participation (at the system level) of approximately one third of the total available multiple sclerosis patient population receiving standard care from each of the participating MS centers in the research collaborative. We will not close the study to participation once the minimum N of 2,000 participants is achieved. Because this is a minimal risk, minimal patient burden study, we will allow "all comers" wishing to participate to do so in the study throughout its entire duration.

c. Define the target population:

The target population for this study is characterized as adults diagnosed with multiple sclerosis who receive standard multiple sclerosis care at one of the four participating multiple sclerosis centers in the research collaborative and meet inclusion and exclusion criteria for study participation.

d. Vulnerable populations

Note: Certain populations are considered vulnerable to coercion and undue influence and are provided with additional protections when participating in a research study.

Identify any of the below populations which you plan to recruit for this study. In addition, complete the form(s) linked with each population as necessary and upload on the 'Supporting Documents' page in Rapport.

☐ Pregnant Women, Fetuses and Neonates
☐ Children
☐ People with impaired decision-making capacity

The following populations may also be considered vulnerable to coercion or other undue influence:

- Prisoners
- People who are economically disadvantaged
- The elderly
- People who are illiterate or do not speak English
- Students and employees

Describe any other potentially vulnerable population(s) and the additional protections provided to them:

Not applicable.

10. Recruitment

Describe method(s) of recruitment. Associated advertisements and other materials to be used for recruitment should be uploaded to the 'Consent Forms and Recruitment Materials' page in Rapport.

Participants will be recruited consecutively at the time of clinic visits at participating MS centers. Research coordinators on site will offer participation by providing the informed consent (IC) form for review. The IC form will be available on the PRO data collection device (see Appendices C & D) which will be on a tablet computer. Paper versions of the IC form will be available for those wishing to have a paper copy. The PRO interface will allow for written informed consent to be completed via tablet computer. Individual level data will be collected for these participants who have given informed consent. Subsequent to enrollment, study staff and the Investigator will verify that participants are eligible per criteria. Upon confirmation of eligibility, participants will be assigned a registration number which will be used to identify data for each participant and link data derived from questionnaires to that abstracted from medical records, as well as to track which patients have given informed consent to participate. Only individual sites will maintain records linking medical record numbers to study identification numbers but these will not be transmitted to the hub site or be used in benchmarking analyses other data analyses for the study. Prior to transmission of aggregated quarterly data to the hub site by participating MS centers, all PHI will be removed save for the study identifier.

Participants will be withdrawn from the study for any one of the following reasons:

- The participant withdraws consent or wishes to discontinue participation.
- The participant is unwilling or unable to comply with the protocol.
- A healthcare provider withdraws the participant from the study for medical reasons.

- Participants will notify the study staff in writing via standard letter or electronic communication of desire to discontinue participation at any time.

11. Informed Consent, Assent, and Authorization

All forms discussed in this section should be uploaded to the 'Consent Forms and Recruitment Materials' page in Rapport

a. Please describe the consent and/or assent process, addressing the following:

- Who will obtain consent/assent from participants
- Where the consent/assent process will take place
- The timeframe for providing information potential participants about a study, having the consent form signed, and beginning study activities
- Any precautions taken to minimize the possibility of coercion or undue influence
- The forms which will be used as well as any aids used to simplify scientific or technical information
- How comprehension will be ensured

Because PHI will be collected in this study, written informed consent will be obtained. Data will be de-identified prior to transmission to the study data analytics center and a highly secure PRO data collection system will be used. Participants will be offered participation and complete informed consent at the time of their regularly scheduled MS care appointments by study staff at participating MS Centers. We anticipate that the risk of PHI dissemination will be minimal and the actual risk of participation for this study is considered minimal and no greater than that of standard multiple sclerosis care. The study principal investigator may also be contacted with study related questions. Patients desiring not to participate may inform local site study staff or the principal investigator at any time to opt out of study participation. Those providing written informed consent will be able to participate in the study.

b. Waiver(s) or alteration(s) may be requested for research that involves no more than minimal risk.

Indicate requested waiver(s) or alteration(s) below. In addition, complete the corresponding section of the Waivers and Alterations Request Form and upload it to the 'Consent Forms and Recruitment Materials' page in Rapport.

☐ For the informed consent *process*
☐ For the *documentation* of informed consent
☐ For the HIPAA Authorization to use and/or disclose PHI
☐ For a waiver of the requirement for medical record documentation

12. Compensation or Gifts

Please describe any payments, gifts or reimbursements participants will receive for taking part in the study:

No remuneration will be provided for participation in this study.

13. Privacy of Participants

Note: Methods used to obtain information about participants may have an effect on privacy. For example:
- Consent discussions or interviews held in public which concern sensitive subjects or behaviors
- Observations of behavior, especially illicit behavior, in quasi-public settings

Describe any activities or interactions which could lead to a breach of privacy and provide a plan to protect participant privacy:

Data collected via the PRO mechanism carries a small risk to privacy because this data includes limited PHI elements. However, because data will be collected using a highly secure PRO mechanism and de-identified and numerically codified prior to transmission to the data analytic center for data analyses, there will be no record or information linking information gathered for

this study to the individuals providing that information. All other data collected for this study will be at the system (MS Center) level and will not link to any form of PHI.

14. Confidentiality of Data

Note: Any person engaged in research collecting information about illegal conduct may apply for a Certificate of Confidentiality from the National Institute of Health.

a. If disclosed, could any of the data collected be considered sensitive, with the potential to damage financial standing, employability, insurability, or reputation?

☒ **No** ☐ **Yes**

If Yes, describe the data or information, the rationale for their collection, and whether a Certificate of Confidentiality will be obtained:

Not applicable.

b. Describe the safeguards employed to secure, share, and maintain data during the study, addressing any of the following which may apply:

- Administrative, i.e. Coding of participant data
- Physical, i.e. Use of locked file cabinets
- Technical, i.e. Encrypted data systems

The data collection procedures in this study include the collection of data with aggregation to the system (MS Center) level and transmission of this data to the data analytics center in de-identified, aggregate form. The PRO platform used for this

study includes sophisticated data security including 128 bit encryption and password protection for data transmissions on database access for study personnel. The use of physical (non-electronic) records for this study will be minimal, and any physical records will be maintained in a double-locked environment, i.e. locked filing system in a locked office. Finally, coding of participant data will be conducted on a system level, i.e. data will be linked to a population of individuals rather than to individual participants, i.e. to one of the four participating MS Centers.

c. Describe the plan for storage or destruction of data upon study completion:

Data for this study will be maintained in de-identified, aggregate format, in a secure, encrypted, and password protected storage system with redundancy back up protection maintained by JCPH, the independent data analytics center. Data will not be maintained by any of the participating sites. After all activities and data analyses are completed for this study, study data will be maintained for the minimum period specified by IRB requirements and then will be permanently deleted.

FIGURES

APPENDICES

References:

1. Institute of Medicine (2001). *Crossing the Quality Chasm: A New Health System for the 21st Century.* Available Online: http://www.iom.edu/~/media/Files/Report%20Files/2001/Crossing-the-Quality-Chasm/Quality%20Chasm%202001%20%20report%20brief.pdf.

2. Institute for Healthcare Improvement (2014). *The IHI Triple Aim.* Available Online: http://www.ihi.org/offerings/Initiatives/TripleAim/Pages/default.aspx.

3. Wennberg J, Gittleson A (1973). Small Area Variations in Health Care Delivery. *Science;* 14(182): 1102-1108. Available: http://www.dartmouthatlas.org/downloads/papers/Science_1973.pdf.

4. Trustees of Dartmouth College (2014). *The Dartmouth Atlas of Health Care* (Website). Available Online: http://www.dartmouthatlas.org/.

5. Welch WP, Miller ME, Welch GH, Fisher E, Wennberg J *(1993).* Geographic variation in physician's expenditures in the United States. *NEJM*; 328:621-7.

6. Fisher E, Wennberg J (2003). Healthcare quality, geographic variations, and the challenge of supply sensitive care. *Perspectives in Biology and Medicine*; 46(1):69-79.

7. U.S. Department of Health and Human services (2013). *The Affordable Care Act* (Website). Available: http://www.hhs.gov/healthcare/rights/law/index.html.

8. Lean Six Sigma (2014). *The Power of Lean Six Sigma* (Website). Available Online: http://engineering.dartmouth.edu/sixsigma/.

9. Trustees of Dartmouth College (2014). *Clinical Microsystems* (Website). Available Online: http://www.clinicalmicrosystem.org/.

10. Northern New England Cardiovascular Network Disease Study Group (2014). *Published Literature* (Website). Available Online: http://www.nnecdsg.org/pub_lit_2.htm.

11. Godfrey MM, Oliver BJ (2014). Accelerating the Rate of Improvement in Cystic Fibrosis Care: Contributions and Insights of the Learning and Leadership Collaborative. *BMJ Quality & Saf*ety; 23:i23-i32 (doi: 10.1136/bmjqs-2014-002804).

12. Marshall BC, Nelson EC (2014). Accelerating implementation of biomedical research advances: Critical elements of a successful 10-year CF Foundation healthcare delivery improvement initiative. *BMJ Quality & Safety Supplement.*

13. Mogayzel P, Dunitz J, Marrow L, Hazle L (2014). Improving chronic care delivery and outcomes: The impact of the cystic fibrosis care center network. *BMJ Quality & Safety Supplement.*

14. Sabadosa KA, Batalden PB (2014). Individuals with cystic fibrosis, families and healthcare professionals: Co-producers of care and quality improvement. *BMJ Quality & Safety Supplement.*

15. Cheng EM, Crandel CJ, Beyer Jr. CT, et al. (2010). Quality Indicators for Multiple Sclerosis. *Multiple Sclerosis*; 16(8): 970–980.

16. American Academy of Neurology (August 2014). Multiple Sclerosis Quality Measures Set (Draft for Public Comment).

17. McDonald KM, Sundaram V, Bravata DM, et al. for the Agency of Healthcare Research and Quality (2007). *Closing the Quality Gap: A Critical Analysis of Quality Improvement Strategies.* Available: http://www.ncbi.nlm.nih.gov/books/NBK44008/figure/A25995/?report=objectonly.

18. Nelson EC, Greenfield S, Hays RD, et al. (1995). Comparing outcomes and charges for patients with acute myocardial infarction in three community hospitals: An approach for assessing value. *International Journal of Quality & Safety in Healthcare*; 7(2):95-108.

19. Nelson EC, Batalden PB, Godfrey MM, Lazar J (2011). *Value by Design: Developing Clinical Microsystems to Achieve Organizational Excellence.* Josey-Bass: San Francisco.

20. Cretin S, Shortell SM, Keelar EB (2004). An evaluation of collaborative interventions to improve chronic illness care. *Evaluation Review,* 28(1), 28-51.

21. de Silva D (2014) for the Health Foundation Evidence Centre (2014). *Improvement Collaboratives in Health Care.*U.K. Health Foundation, London.

22. Kilo CM (1998). A framework for collaborative improvement: Lessons from the Institute for Healthcare Improvement's Breakthrough Series. *Quality Management in Health Care*, 6(4), 1-13.

23. Godfrey MM (2013). *Improvement Capability at the Front Lines of Healthcare: Helping through leading and coaching* (PhD Dissertation & Publications). Available Online: http://hj.diva-portal.org/smash/get/diva2:640804/FULLTEXT01.pdf

24. Godfrey MM (2013). *Improvement Capability at the Front Lines of Healthcare: Helping through leading and coaching* (PhD Dissertation & Publications). Available Online: http://hj.diva-portal.org/smash/get/diva2:640804/FULLTEXT01.pdf.

25. Godfrey MM, Andersson-Gare B, Nelson EC, Nilsson M, Ahlstrom G (2013, in Press). Coaching Interprofessional Health Care Improvement Teams: The Coachee, the Coach and Leader perspectives. *Journal of Nursing Management*. Doi: 10.1111/jonm.12068. [Epub ahead of print]

26. Grumback K, Bainbridge E, Bodenheimer T (2012). Facilitating improvement in primary care: The promise of practice coaching. *The Commonwealth Fund*, 1605(15), 1-14.

27. Harvey G, Loftus-Hills A, Rycroft-Malone J, et al. (2002). Getting evidence into practice: The role and function of facilitation. *Journal of Advanced Nursing,* 37(6), 577-588.

28. Gustafson DH, Quanbeck AR, Robinson JM, et al. (2013). Which elements of improvement collaboratives are most effective? *Addiction.* DOI: 10.1111/add.12117

29. Rapkin BD, Weiss ES, Lounsbury DW, et al. (2012). Using the interactive systems framework to support a quality improvement approach to dissemination of evidence-based strategies to promote early detection of breast cancer: Planning a comprehensive dynamic trial. *Am J Community Psychol.* doi 10.1007/s10464-012-9518-6.

30. West SG, Duan N, Pequegnat W, et al. (2008). Alternatives to the Randomized Controlled Trial. *Am J Public Health;* 98(8): 1359-1366.

31. Huang X, Rosenthal MB (2014). Transforming Specialty Practice — The Patient-Centered Medical Neighborhood. *NEJM;* 370(15): 1376-78.

32. Coleman K, Austin BT, Brach C, Wagner EH (2009). Evidence On The Chronic Care Model In The New Millennium. *Health Affairs;* 28(1): 75-85.

33. Sia C, Tonniges TF, Osterhus E, Taba S (2004). History of the Medical Home Concept. *Pediatrics:* 113:1473-78.

34. American College of Physicians (2010*). The Patient-Centered Medical Home Neighbor: The Interface of the Patient-Centered Medical Home with Specialty/Subspecialty Practices*. Philadelphia: American College of Physicians.

35. Institute for Healthcare Improvement (2014). *The Plan-Do-Study-Act Cycle.* Available Online: http://www.ihi.org/knowledge/Pages/Tools/PlanDoStudyActWorksheet.aspx.

36. Weinfurt, K. P. (2000). Repeated Measures Analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and Understanding Multivariate Statistics* (pp. 317–361). Washington, DC: American Psychological Association.

37. Weissman NW, Allison JJ, Kiefe CI, et al. (1999). Available benchmarks of care: The ABCs™ of benchmarking. *Journal of Evaluation in Clinical Practice*;5:269-281. Available Online: http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=5185359&site=ehost-live&scope=site.

38. Benneyan JC, Lloyd RC, Plesk PE (2003). Statistical process control as a tool for research and healthcare improvement. *Quality and Safety in Health Care*;12: 458-464. Available Online:http://qshc.bmj.com/cgi/reprint/12/6/458.

39. Thor J, Lundberg J, Ask J, et al. (2007). Application of statistical process control in healthcare improvement: A systematic review. *Quality and Safety in Health Care*;16(5): 387-99. Available Online: http://qshc.bmj.com/cgi/reprint/16/5/387.

40. Johnson K, Brooks BR, Cohen JA, Ford CC, Goldstein J, et al. Copolymer 1 reduces relapse rate and improves disability in relapsing-remitting multiple sclerosis: Results of a phase III multi-center, double-blinded, placebo-controlled trial. Neurology. 1995; 45:1268-1276.

41. Panitch H, Goodin DS, Francis G, Chang P, Coyle PK, O'Connor P, et al. Randomized, comparative study of interferon beta-1a treatment regimens in MS: The EVIDENCE trial. Neurology. 2002; 59(10):1496-1506.

42. Goodin D, Arnason B, Coyle P, et al. The use of mitoxantrone (novantrone) for the treatment of multiple sclerosis. Neurology. 2003; 61:1332-1338.

43. Kargiotis O, Paschali A, Messinis L, Papathanasopoulos P. Quality of life in multiple sclerosis: Effects of current treatment options. Int Rev Psychiatry. 2010; 22(1):67-82.

44. Trojano M, Pellegrini F, Paolicelli D, et al. Real-life impact of early interferon beta therapy in relapsing multiple sclerosis. Annals of Neurology. 2009; 66(4):513-520.

45. Bell C, Graham J, Earnshaw S, Oleen-Burkey M, Castelli-Haley J, Johnson K. Cost-effectiveness of four immunomodulatory therapies for relapsing-remitting multiple sclerosis: A Markov model based on long-term clinical data. Journal of Managed Care Pharmacy. 2007; 13(3):245-261.

46. Barkhof F, Simon JH, Fazekas F, et al. (2012). MRI monitoring of immunomodulation in relapse-onset multiple trials. *Nature Reviews Neurology*; 8:13-21.

47. Ogrinc G, Headrick LA, Morrison LJ, Foster T. Teaching and assessing resident competence in practice-based learning and improvement. J Gen Intern Med. 2004;19(5, pt 2):496–500.