University of Alabama at Birmingham

UAB Digital Commons

All ETDs from UAB

UAB Theses & Dissertations

2021

Applications of Longitudinal Machine Learning Methods in Multi-Study Alzheimer's Disease Datasets

Charles F. Murchison
University of Alabama at Birmingham

Follow this and additional works at: https://digitalcommons.library.uab.edu/etd-collection

Part of the Public Health Commons

Recommended Citation

Murchison, Charles F., "Applications of Longitudinal Machine Learning Methods in Multi-Study Alzheimer's Disease Datasets" (2021). *All ETDs from UAB*. 545.
https://digitalcommons.library.uab.edu/etd-collection/545

This content has been accepted for inclusion by an authorized administrator of the UAB Digital Commons, and is provided as a free open access item. All inquiries regarding this item or the UAB Digital Commons should be directed to the UAB Libraries Office of Scholarly Communication.

APPLICATIONS OF LONGITUDINAL MACHINE LEARNING METHODS IN MULTI-STUDY ALZHEIMER'S DISEASE DATASETS

by

CHARLES F. MURCHISON

JEFF M. SZYCHOWSKI, COMMITTEE CHAIR
GARY R. CUTTER
BYRON C. JAEGER
RICHARD E. KENNEDY
ERIK D. ROBERSON

A DISSERTATION

Submitted to the graduate faculty of The University of Alabama at Birmingham,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

BIRMINGHAM, ALABAMA

2021

Copyright by
Charles F. Murchison
2021

APPLICATIONS OF LONGITUDINAL MACHINE LEARNING METHODS
IN MULTI-STUDY ALZHEIMER'S DISEASE DATASETS

CHARLES F. MURCHISON

BIOSTATISTICS

ABSTRACT

Advances in statistical learning models for prediction have led to broader application across a variety of disciplines, granting generalizations and adaptations that were previously intractable even with advanced computational techniques. Among these is the allowance of correlated data with inherent paneled structure such as longitudinal or clustered data; adjustments which have already begun to be applied to a variety of supervised and unsupervised machine learning methods which had previously focused on cross-sectional data. These modifications have seen rudimentary testing in a number of applied disciplines where correlated data is commonly observed, including medical and clinical research. One field in particular that has garnered interest is Alzheimer's disease and related dementias. As this disorder is characterized by a prolonged and progressive disease course with an extensive variety of potential biomarkers, its feature-dense datasets with repeated patient measures are well suited for applications of machine learning prediction while utilizing longitudinal modifications. While some novel adaptations of longitudinal machine learning methods have already been tested in the realm of Alzheimer's disease, there has not yet been a comprehensive evaluation to compare these techniques against each other or against widely accepted standards such as traditional inferential techniques like mixed-effects regression. Nor has there been rigorous investigation into how subject-specific effects can impact the error and bias of these predictions and the distinctions which may arise when developing entire temporal profiles as compared to the forecasting of future observations while leveraging previously observed data. This dissertation addresses these deficiencies in the literature by directly comparing a variety of machine learning techniques with longitudinal adaptations against each other and reference standards using a large, multi-study Alzheimer's disease meta-database as well as assessing the role of subject-specific effects using synthetic data. This study is especially comprehensive, considering both continuous and categorical outcomes as well as differences when generating whole profiles *de novo* or forecasting of future observations based on prior sequences. With its emphasis on longitudinal data, this study considers not only predictive capacity for unobserved data using population-level characteristics, but also prediction of future observations using a variety of subject-specific effects.

Keywords: machine learning, prediction, longitudinal, subject-specific effects, Alzheimer's disease

DEDICATION

This dissertation is dedicated to my parents, Dick and Ann Murchison. They will forever remain the stalwart pillars of my life and I would not be where or who I am today without their love and support.

ACKNOWLEDGEMENTS

An enormous thank you to my dissertation committee whose counsel and insight have been instrumental to the success of this study.

Dr. Jeff Szychowski: In addition to being my chair, Jeff has been my academic advisor throughout my time at UAB which I complicated in many ways. I can also never thank him enough for the belief and confidence he had in me to let me craft my project in the way that I saw fit and tailor it to what I wanted to see. It takes an exceptional amount of trust for an advisor or chair to be willing to show that level of faith in a student and for that I will always be appreciative.

Dr. Richard Kennedy: I very much consider myself Richard's padawan. The consistent generosity he has shown in his time, knowledge and experience has benefited me immensely and I would not have had a fraction of my success here without his guidance. There is no other person I have learned as much from during my time at UAB and he remains the example of what I strive to be in my own research endeavors.

Dr. Byron Jaeger: Byron is the statistician I wish I had been ten years ago, and his expertise and insight have supported many of the more technical aspects of this project's implementations. More importantly, he was especially supportive during the nadir of this work, helping me have a better appreciation of what this dissertation should be and marking a turning point for me and this study.

Dr. Erik Roberson: Also serving as my supervisor, I have long known Erik is a remarkable clinician and research. However, I believe his true calling is mentorship. You only need to see the caliber of students and trainees who have come through his lab and the extraordinary pride he takes in watching their success and growth to share this understanding. I will be forever grateful to have benefited from even a portion of his counsel.

Dr. Gary Cutter: There is one individual responsible for my arrival in Birmingham, and it most certainly is not me. I came to Gary with a sob story and in response he gave me an opportunity. He has been by my side at every step of this journey, and I am truly gratified that he has been able to be here all the way to the end.

Finally, a thank you to my colleagues and collaborators at the Alzheimer's Disease Research Center and Center for Neurodegeneration and Experimental Therapeutics as well as the faculty, staff, and students in the department of Biostatistics at the University of Alabama at Birmingham. Every one of you have made this an incredible experience.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table *Page*

LIST OF FIGURES

Table *Page*

LIST OF ABBREVIATIONS

1D CNN	one-dimensional convolutional neural network
AD	Alzheimer’s disease
ADAS	Alzheimer’s Disease Assessment Scale
ADAS-Cog	Alzheimer’s Disease Assessment Scale – Cognitive Subscale
ADCS	Alzheimer’s Disease Cooperative Study
ADNI	Alzheimer’s Disease Neuroimaging Initiative
ADRD	Alzheimer’s disease and related dementias
AE	absolute error
ANN	artificial neural network
APOE4	apolipoprotein E4
AUC	area under the curve
AV	absolute value
AVB	absolute value of the bias
BR	beta regression
CART	classification and regression tree
CDR	Clinical Dementia Rating
CI	confidence interval
CNN	convolutional neural network
CPath	Critical Paths in Alzheimer’s Disease

DN BR	*de novo* beta regression
FNN	feed-forward neural network
GED	general equivalency degree
GLMM	generalized linear mixed model
HS	high school
LASSO	least absolute shrinkage and selection operator
LSTM	long short-term memory
LSTM RNN	long short-term memory recurrent neural network
MAE	mean absolute error
MCI	mild cognitive impairment
MERF	mixed-effects random forest
ML	machine learning
MMSE	Mini-Mental State Exam
MoCA	Montreal Cognitive Assessment
MRI	magnetic resonance imaging
NN	neural network
NRI	net reclassification improvement
PET	position emission tomography
PLE	population-level effect
ReLU	rectified linear unit
RMSE	root mean square error
RNN	recurrent neural network
ROC	receiver operator characteristic

ROC AUC	area under the receiver operator characteristic curve
SMAE	symmetric mean absolute error
SSE	subject-specific effect
SVM	support vector machine
VC	Vapnik-Chervonenkis

INTRODUCTION

Prediction versus Inference in Statistical Learning

As a discipline, statistical methods have largely been cast as serving one of two primary roles: either to draw inferences about the association between covariates of interest and a response outcome, or prediction of a response given certain parameterizations and corresponding covariate sets. The former in particular is frequently leveraged by biostatisticians and their collaborative researchers, enabling a better understanding of associations between observed variables and outcomes while facilitating formal hypothesis testing. However, interest in the role prediction, through statistical learning, can play as a companion to traditional inference methods has greatly increased in the last several decades. Much of this impetus stems directly from advances in technology and improvements in computational power which has enabled many of the heuristic and algorithmic approaches central to machine learning which were previously intractable. These improvements have corresponded with greater capacity to collect, refine, and develop large-scale datasets, and their combination has led to substantial growth in predictive learning methods which have seen implementations that would otherwise never have been possible. Many of the predictive models and algorithms which have been developed have fallen under the banner of machine or statistical learning and now comprise an entire field in their own right.

Machine learning, in general, has placed less emphasis on the interpretation of covariate associations and inference and instead focused on the ability to identify underlying features through dimensional reduction or to either accurately predict a response using classification and labelling or to predict numeric real-valued outcomes akin to regression. Many of the goals and concerns central to inferential methods, such as minimizing bias, are given secondary importance, with high accuracy of future predictions generally considered paramount. In addition, the concept of hypothesis testing is largely left by the wayside even though it is a frequent goal of inferential statistics specifically. This difference in prioritization has let machine learning, as a separate collection of techniques, be accepted as another facet of statistics as it has attained more widespread popularity and been utilized in several different applications. Notable advances have been most widely seen in domains where large corpuses of data are common, such as image annotation, natural language tasks, and high-dimensional dataset processing. Predictive learning methods have also evoked notable interest in several fields of medical research. Recognizing their readily available and sizable datasets, medical and clinical disciplines have been among the most prominent spheres of research attempting to leverage the techniques of machine learning against the vast data-rich resources available to them such as electronic health records and high-resolution medical images. While inference and interpretation of variable associations will continue to be a central aspect of statistical research and application, predictive machine learning techniques have already demonstrated their capacity to support and buttress the classical inference methods many often think of in prototypical statistical analysis.

Responses and Supervision for Machine Learning Outcomes

As various algorithms have developed, it has been necessary to delineate the several methods of statistical learning akin to the same way different inferential methods are distinguished. Much like inferential linear regression models, machine learning methods are often defined by the nature of their response variables. Ordinary least-squares models for normally distributed scale variables and logistic regression for binomially distributed outcomes have their own machine learning analogues including regression and binary classification. Similarly, extensions such as multinomial classification for categorical variables with more than two levels can be applied to both inferential models and predictive designs. Even further augmentations seen in inferential models, like splines to allow for non-linear relationships, have corresponding methods for learning and prediction such as generalized additive models which are also capable of characterizing non-linear effects. As so much of statistical inference has informed predictive methods, it comes as no surprise many of the unique facets of response outcomes in standard statistical and biostatistical techniques have corollaries also seen in machine learning implementations.

However, machine learning also has utility even in the absence of expressly defined response variables. This in turn leads to a design consideration less common in standard statistical methods with the notion of supervision, namely the amount of prior knowledge on the outcome of interest. In machine learning, there is often a distinction between *supervised* learning and *unsupervised* learning. Generally speaking, in supervised learning the label of the output, whether numeric or categorical, is known prior to the development of a predictive model, and these known outputs directly inform the structure and training methods of the input features to forecast response predictions. In unsupervised learning, the outputs are either unknown or unobservable and relationships and feature associations are defined solely by the underlying structure of the data inputs without any sort of oversight or consideration of the output. This scenario is actually quite common and there are many situations in which the outputs may be unknown. For example, human annotation of the data, the general gold-standard for curation, may be too resource intensive and as such is intractable. This might be the case for a collection of hundreds of thousands of pictures being used for image recognition. Similarly, a researcher may want to take a completely *de novo* approach to data clustering with a desire to use unsupervised methods to generate hypotheses for later inferential hypothesis testing.

Notably, unlabeled data is not solely utilized by unsupervised machine learning techniques. While unsupervised methods are generally focused on clustering of data or reduction of dimensionality of dataset features, prediction tasks may still be desired even in the absence of curated outputs. In these situations, the machine learning model is tasked with teaching itself labels without any sort of reference based on prior curation, a method referred to as *self-supervision*. Self-supervision learning has become very popular in a variety of machine learning applications, especially in image processing where feature identification or classification is desired but labelling of images is not available. Many of the advances in deep learning neural networks and autoencoders have applied these self-supervision approaches to their learning tasks. A common example is the denoising of images where reconstruction of an otherwise unknown original image is the desired output when the input is solely based on a variety of perturbed versions of the image. Although self-supervision methods have been frequently developed for use with unlabeled data, their goal of classification can still be utilized with annotated datasets to reinforce and support model training. As such, there are many circumstances where unsupervised and self-supervised methods on unlabeled data can be expressly purposed and these methodologies, by their very nature, require special consideration compared to prototypical supervised techniques.

It should be noted that data supervision is not a strictly binary consideration but rather a spectrum wherein fully curated data or data without any labelled outcome are the two extremes of supervised and unsupervised, respectively. There are several methods that lie in between within the realm of *semi-supervised* methods where partial annotation is available. This is frequently seen when human labelling of a full dataset is especially expensive in either time or resources but labelling of a smaller subset is much more feasible. In these situations, model training can make partial utilization of the known data to help direct and inform the associations which subsequently arise from the unlabeled outcomes. Furthermore, data in absence of an outcome is also not unheard of in inferential statistics. For example, log-linear models commonly seen in systems analysis can also be created in the absence of a strict output but can still identify associations among features and covariates and provide the corresponding effect sizes and strengths of associations in the forms of confidence intervals and p-values. However, the idea of model supervision and the utilization of unlabeled data is much more common in machine learning, especially when applied to large corpuses of data with thousands or millions of features.

Correlated and Longitudinal Structure in Response Measures

Beyond similarities in the nature of their responses and outcomes, machine learning algorithms also have a marked overlap with traditional inference methods in several of their necessary underlying assumptions. Much in the same way they are necessary for valid inference in relationships, certain fundamental assumptions are required to increase accuracy and increase prediction veracity in statistical learning. One assumption shared by the more basic forms of both inferential and predictive methods is the independence of observations. When the data has natural structure, for example through correlated or repeated measures, the assumption of independence is violated, and model adjustments are required regardless of the ultimate goal of the statistical design. In inferential methods this can be handled through techniques such as repeated measures analysis of variance, generalized linear regression, or mixed-effects models. In these cases, additional structure is applied to the residual variance of a model while also adjusting the ways in which models can vary, generally through redefining the degrees of freedom. Machine learning algorithms are no less susceptible to deviations in prediction that are dependent on any inherent correlation of the data, whether in the known response of a supervised method or *a priori* clustering of input features prior to unsupervised associative models.

Akin to how inferential methods began with assumptions of independence before generalizing to correlated data, machine learning implementations first began with cross-sectional data, establishing the fundamental groundwork before attempting to extend to more complex and refined data designs. These considerations of correlated data have taken multiple forms in the machine learning literature. Some have obvious analogs to inferential methods, such as accounting for repeated measures taken longitudinally and leveraging the knowledge of serial correlation to aid in within-unit predictions. Others are more esoteric, such as using known language structures to aid in natural language processing or proximity of image sub-units which display increased feature similarity with decreasing distance. Not only has this led to several extensions of previously developed statistical learning techniques with cross-sectional origins, for example applying random effects components much like mixed-effects models, these considerations of structured data have led to entirely new techniques which have supplanted their original designs, most notably with advances in neural networks and other self-supervised methods. Regardless, there has been increased appreciation for the role correlated data plays in predictive statistical learning and the impact it can have on accuracy and validity cannot be overstated.

Even outside of machine learning paradigms, other considerations of correlated data on prediction are important. For example, subject-specific effects can be utilized in a number of ways including complete suppression in order to rely solely on population-level effects, imputation of probable subject-specific covariate values based on prior model parameters such as the covariance matrices of mixed-effects models, or leveraged directly as known subject-specific values based on previously observed values when models are built *de novo*. The exact behavior of these subject-specific effects is not especially well known even in standard inferential statistical models, let alone for machine learning methods. The role these subject-specific effects have could very easily be context dependent with different designs having specific strengths depending on how they are utilized, such as when calculating cohort level effects on average as opposed to specific predictions for an individual. In addition, when model parameterizations are provided instead of calculated directly, there may be severe consequences with differences in prediction capacity even for the same type of model. This is just another unique aspect of longitudinal data that is critical to remember both for standard inferential methods as well as machine learning designs.

Applications in Medical Research

Many of the considerations and extensions in statistical learning have been directly informed by the attempts to apply the methodologies in real-world disciplines and scenarios. As mentioned, one of the most prominent of these fields ripe for application is in clinical medical research. At its most basic level, clinical research meets many of the various criteria previously discussed. Correlated data is exceedingly common, with some studies focusing on the patient as the fundamental unit with several repeated in-patient measures while others consider natural clustering of sets of patients due to similarities in demographics, diagnoses, or dispositions. In both cases, correlation within dataset panels has severe repercussions on final prediction if proper allowances are not made. Furthermore, this gives an additional enhancement on how prediction of correlated data can be utilized: not only can a new, previously unobserved unit or patient be generated based on a specific set of features or characteristics, but a subsequent measure for a patient who has already been previously observed can also be forecast. In this latter case not only are the population-level features such as demographics and patient characteristics utilized, but the unique subject-specific adjustments are also applied, leveraging the additional sequence information inherent in the serially correlated data. This can ostensibly give additional validity to any within-unit forecast response, potentially yielding even greater levels of accuracy than would otherwise be seen by only using population-level features.

As previously mentioned, the complexion of clinical and medical data readily lends itself towards machine learning applications by its very nature. High dimensionality is especially common, with feature sets often greatly outstripping the number of discrete data units (e.g. patients) in sheer scope and breadth. Assays on hundreds of biomarkers from blood, spinal fluid, and other vectors are commonly taken in the course of both clinical practice as well as medical research. The advances in assays and informatics analyses have also created enormous corpuses of "omics" data. Patient populations can now be characterized across metrics which number from hundreds of gene alleles or protein products, to thousands of single-nucleotide polymorphisms, to even larger data corpuses such evaluations on the near countless species within the gut microbiome. The width of these ever-increasing datasets requires careful handling and standard statistical inferential techniques often perform poorly in these scenarios due to overfitting and multiplicity of hypotheses. However, dimensional reduction and feature collapse are hallmarks of statistical learning and extraction of underlying latent structures can still lead to high-quality prediction even when inference of association may be limited.

Another application of statistical learning to medical research which warrants mentioning is imaging. In much the same way computational advances have contributed to machine learning, medical imaging has greatly benefited from its own technological advances. This has led not only to images of higher resolution and quality, but whole new modalities of imaging and extensions from two-dimensional images comprised of pixels to three-dimensional representations built from voxels. Accompanying these imaging methods is the desire to assist manual curation, and even apply automation, which have spearheaded unique applications of machine learning in the domain of image processing.

These include feature identification within individual images as well as larger-scale classification and categorization tasks for entire image collections. In turn, this has led to a variety of machine learning applications that have seemingly larger departures from the underlying statistical techniques that preceded them; however, the fundamentals of the classification and evaluation still apply in many respects to the original inferential and statistical learning methods which preceded them.

Specific Applications for Alzheimer's Disease and Related Dementias

One set of disorders is of particular interest in the machine learning field for many of the reasons previously discussed: Alzheimer's disease and related dementias (ADRD). Alzheimer's disease (AD) is one of the most prominent medical concerns both nationally and across the globe. The aging population, particularly within the western world, has made the prevalence of this disease reach all-time highs and is one of the leading causes of death among those over the age of 65. Rates from 2018 estimated that over 5.7 million patients are living with AD in the United States alone and this is expected to increase to 13.8 million patients by 2050. Accompanying this disease is a burden of direct care, time investment and material cost for patients, caregivers, doctors, and taxpayers; an estimated $277 billion dollars was spent on AD and similar dementias in the United States in 2018, with $186 billion being directly paid by Medicare. This combination of cost and pervasiveness has been accompanied by a necessity to better identify and treat patients with AD and placed particular emphasis on predicting future changes in cognition, whether as direct cognitive outcomes used as diagnostic criteria or predicting future cognitive status. One of the unique characteristics of AD is its prolonged time course with many of the pathological and biological hallmarks occurring years before clinical symptoms begin to manifest. This has led to a corresponding increase in research into the ability to foresee future changes in cognition and dementia status in patients based on these markers. The ability to accurately predict cognitive trajectories and changes in patients can in turn help direct interventional treatments and ideally even aid in prevention while aiding researchers when designing clinical trials.

As mentioned, ADRD research bears many hallmarks and prerequisites of statistical learning. This includes the nature of its high-dimensional datasets with extensive neuroimaging, genetic, and biomarker profiling which have been collected to characterize the disease more fully. These covariate-rich datasets can be used as labeled data in a supervised fashion to identify feature sets which can either predict metrics used in AD research or to classify patients wholesale based on cognitive and functional status. Unsupervised and semi-supervised applications are also valid as definitions of AD can vary based on cognitive capacity, clinical function, and expression of specific pathologies and biomarkers. In addition, the progressive nature of the disease and its protracted time course have led to long-term studies comprising sets of collected data within patients who are followed for years and even decades. Thus, not only can novel predictions of unobserved or hypothetical patients be considered, but so can future predictions of patients within datasets with their unique, individual-level factors used to aid prediction. Taken in combination, this has made ADRD research a prime candidate for the applications of machine learning techniques. The potential in ADRD has been recognized for some time although there have been several deficiencies in approaches which seek to be addressed.

Longitudinal Machine Learning and Alzheimer's Disease

Although there have been several studies which have attempted to use machine learning paradigms to predict cognitive status and neuropsychological outcomes, nearly all have been focused on cross-sectional data, merely drawing associative conclusions rather than directly predicting longitudinal trajectories and change within patients and subjects, or compressing longitudinal measures into single aggregate values like annualized change. This has limited the utility and predictive capacity of these machine learning methods since they have been unable to make use of the longitudinal nature of AD datasets which collect data repeatedly over time, whether as part of formal interventional trials, observational studies, or during the data collection process in the course of standard clinical care practice. While some domains such as natural language processing have more stringently required the ability to leverage temporal or sequential context in training and prediction, it is only recently that machine learning modalities applied in other disciplines such as medical research have initiated investigation as to how to best make use of time series and sequence data. Fortunately, this has begun to lead to the generalization of many classic statistical learning applications to now account for panel and longitudinal data, whether through use of mixed-effect analogs to model subject or cluster-specific effects, or self-referential memory designs that retain prior information during training to inform future responses.

As these methods have been developed and refined in other areas of medical research, they have started to be used to aid in the prediction of Alzheimer's disease diagnoses and changes in cognition. However, methods presented within the literature are often considered in isolation, with novel algorithms presented and benchmarked using synthetic data in conjunction with widely available AD datasets such as the long-standing Alzheimer's Disease Neuroimaging Initiative (ADNI). Unfortunately, these longitudinal methods are largely evaluated with only rudimentary metrics on single types of response outputs with minimal comparisons between methods. As such, there has not yet been a broader consensus on the potentially situational benefit of these methods where they are comprehensively assessed to ascertain the role they will ultimately have in cognitive decline research and clinical practice, especially when compared to long-standing models which are already accepted by the AD research community. Different outcomes serving different research and clinical goals, generalization across multiple studies, and comparison of more disparate learning designs are simply unavailable which limits the full consideration of these predictive methods and the field of neurodegeneration suffers because of this.

This dissertation bridges the deficiency between the initial implementation of these longitudinal machine learning techniques and their potential application in AD research by extensively evaluating and comparing these more novel methods both against each other as well as against commonly accepted and pre-specified standards using the more traditional inferential modelling paradigms. This enables more direct comparison of the roles both the heavily investigated supervised and discriminative techniques as well as the more novel self-supervised and generative deep learning methods can serve in the study of Alzheimer's disease. Furthermore, different outcomes providing different contextual purposes are considered, specifically neuropsychological scores of cognitive ability frequently used as clinical research outcomes as well as classification of cognitive status inherent to standard clinical practice. Special examination is also given to the different types of longitudinal prediction provided by these models when making direct use of prior observations. Specifically, models are evaluated not only for their ability to predict whole trajectories of outcomes for previously unobserved patients solely using population-level characteristics but also for their predictive capacity to forecast future observations within previously observed units by leveraging individual-specific model adjustments to outcomes. Taken together, this wide-ranging evaluation more fully characterizes the impact of longitudinal machine learning in predicting cognitive changes and help inform its utility in the field of Alzheimer's disease research as a whole.

OBJECTIVES, AIMS, AND IMPORTANCE

The primary objectives of this dissertation research were to compare the predictive ability of a variety of longitudinal machine learning techniques when specifically applied to a harmonized dataset of subjects with varying levels of cognitive and functional impairment, ranging from the cognitively intact to patients with known diagnoses of Alzheimer's disease. These patients were assembled from a datastore of various clinical trials and observational studies (the AD meta-database) with an emphasis on repeated measures of cognitive outcomes and feature sets comprising both persistent baseline values as well as time-dependent variables. Two types of outcomes were considered based on class of the variables and their contextual importance. The first was the Alzheimer's Disease Assessment Scale – Cognitive Subscale (ADAS-Cog), a metric commonly used to evaluate cognitive impairment in research studies and interventional trials, as a continuous scale outcome using regression-based designs. The other was classification of cognitive status, specifically comparing cognitively intact to cognitively impaired, with the binary variable built based on the staging score of the Clinical Dementia Rating (CDR) with scores of 0 for normal patients compared to those with scores of 0.5 or greater indicating any level of cognitive impairment. These two outcomes were selected as contextual and timeframe counterpoints with the ADAS-Cog used as a measure of later stage dementia with particular utility as a metric in AD research while changes in cognitive status emphasized an earlier phase of the disease with specific benefit for patients seen in a clinical setting. Additionally, predictive capacity of the models was compared based on their ability to develop an entire whole-subject trajectory or temporal profile of an individual using population-level features as well as their performance to forecast a future observation in a previously observed participant by leveraging subject-specific adjustments to outcomes alongside the population-level characteristics. In addition to direct comparison of these models, these data-driven techniques were also compared against commonly used standards of regression and classification using more traditional inferential statistical methods. Specifically, linear mixed-effect models were developed for the ADAS-Cog (regression on a normalized beta-distributed response) using pre-specified parameterizations from the widely leveraged Critical Paths for Alzheimer's disease (CPath) AD simulation software and cognitive status built with ad hoc models (logistic mixed-effects regression) using a portion of the primary meta-database. Finally, special consideration was given to the exact contribution of subject-specific effects for these different designs, comparing prediction performance under a variety of scenarios including complete suppression of subject-specific effects, their imputation from model parameters, and directly leveraging previously observed sequences and known fitted values. This comprehensive evaluation was designed to give a better understanding of how these various longitudinal statistical learning methods could support AD research.

The first aim was to evaluate and compare supervised and discriminative longitudinal machine learning techniques for regression and classification in the multi-study Alzheimer's disease dataset. This aim focused on the more commonly implemented discriminative models for machine learning with adaptations to facilitate their use on longitudinal data. These methods initially focused on three primary classes of statistical learning techniques: penalized or regularized regression (ℓ1 LASSO, ℓ2 ridge regression, and elastic net on mixed-effects regression models), ensemble methods using decision trees (bagging and boosting on mixed-effects regression trees, and mixed-effects random forest), and support vector machines using multiple kernels to account for long-term and short-term longitudinal effects. Over the course of the study, emphasis was eventually placed on the ensemble methods as the most robust and well characterized of these supervised designs.

The second aim sought to evaluate and compare self-supervised or generative deep learning techniques for regression and classification in the same multi-study Alzheimer's disease cohort. This aim was structured around the more novel deep learning neural networks that have seen wider use in fields of image processing and natural language tasks and apply their time-dependent extensions to the longitudinal data collected in the AD meta-database. These methods have demonstrated straightforward transfers to sequence-based data making them well-suited to longitudinal data representations. The methods under consideration were long short-term memory (LSTM) recurrent neural networks (RNN) and one-dimensional convolutional neural networks (1D CNN) for time series data. These designs were explicitly selected as they both have training paradigms that naturally translate to longitudinal data applications.

The third aim moved away from model-level evaluations to directly contrast the role of subject-specific effects on prediction when used in a variety of model designs and how those effects differed according to model parameterization, type of subject-specific effect, and their application when generating whole temporal trajectories for subjects or when forecasting future observations. Investigation of subject-specific effects on generating whole outcome trajectories and observation forecasting compared suppression of subject-specific effects to only use population-level covariates, imputation of subject-specific effects based on model parameterizations, and directly leveraging subject-specific effects when forecasting future observations based on prior data used during model building. Superiority in regression-based prediction metrics for the ADAS-Cog was compared among the pre-parameterized CPath reference model, a mixed-effects regression model with an equivalent structure but a *de novo* parameterization, and a supervised machine learning ensemble method, all within and across subject-specific effect designs.

Overall, the purpose of this dissertation was to expand upon the field of statistical and machine learning to better understand the role it can play in aiding with the prediction of cognitive outcomes in the field of Alzheimer's disease and related dementias. Of note, this research was designed to fill a deficiency in the literature by more comprehensively characterizing longitudinal machine learning methods when applied to AD and mild cognitive impairment (MCI) research datasets. In addition, it was developed with a goal of better understanding the impact the repeated measures and correlated data commonly collected in AD studies have on these learning methods and their predictive ability. Additionally, this dissertation was applied to a real-world dataset developed across multiple studies and harmonized, mirroring the increasingly common trend of combining and aggregating multiple distinct datasets into unified data repositories. Such research designs are consistent with the goals of ADRD research as earlier and earlier interventions are investigated to postpone the onset of cognitive decline and dementia. This comprehensive evaluation not only compared these various discriminative and generative machine learning models against each other, but also against widely accepted referential standards which had been developed using more traditional inferential approaches to prediction, simulation, and data generation.

Of special interest was the predictive capacity of the machine learning models in question when utilized for both entire time course trajectories of unobserved data units as well as forecasting future observations when leveraging previously observed data for model generation. The unique nature of longitudinal data allows for two types of prediction: either generation of new data profiles wholesale or extending previously observed data to forecast subsequent measures. As such, special care was taken to explicitly consider both types of longitudinal prediction which is frequently an underappreciated aspect of time-series data. This gave a unique comparison point to identify the utility these models could provide as predictive tools with the idea that some designs would be better suited for novel data generation while others may display increased performance to extend observations within a dataset. This also allowed for a better understanding of how machine learning prediction, as a tool, could be leveraged in AD research and which model designs were better suited for predicting data *de novo* or as a unit/patient-level forecasting tool. Additionally, this directly tied to the notion of the impact subject-specific effects can have on these predictions and whether performance differed between trajectory generation and observational forecasting when subject-specific effects were suppressed, imputed, or directly leveraged.

Finally, this dissertation provided a unique opportunity to investigate the role machine learning can play in AD/MCI research more deeply in the context of both research studies, through the use of the ADAS-Cog, and clinical practice, via the CDR-Based cognitive status classification. Of particular benefit and importance is an understanding of the impact longitudinal considerations play in the training, testing and validation of statistical learning regressors and classifiers. This is especially poignant as electronic health record mining, large-scale datasets, and multi-study cohorts are developed in both clinical practice and interventional research. As the machine learning field continues to make use of different modalities, the results of this research can help contribute to the understanding discriminative and generative models of learning have in predicting AD outcomes at a variety of timepoints, contexts, and scenarios. Taken together, this study provides an excellent analysis well poised to aid in the overall understanding the various characteristics of longitudinal machine learning can play in predicting cognitive outcomes and facilitating its related research.

LITERATURE REVIEW

Initial Beginnings

Machine and statistical learning, as a concept, is not especially novel. In fact, the first published instance of an algorithm designed to be executed on a machine was presented by Ada Lovelace in 1843. Designed to calculate a set of Bernoulli numbers on a proposed general-purpose computer called the Analytical Engine, this is widely recognized as the first computer program (Fuegi & Francis, 2003). Although Lovelace's algorithm laid the groundwork and additional mathematical models of learning would continue to be designed, any sort of practical realization of a machine learning system was still over a century away. It would not be practically realized until 1958 when Frank Rosenblatt developed the perceptron algorithm and the associated Mark I Perceptron machine at the United States Office of Naval Research (Rosenblatt, 1958). The perceptron, unsurprisingly, was somewhat rudimentary in design; it was only capable of learning linearly separable patterns and limited in its classification capacity. For example, single-layer perceptrons like Rosenblatt's are incapable of solving exclusive or (XOR) level problems (Minsky & Papert, 1969). Although limited in scope, several advances in implementation of machine learning designs were carried out, with learning programs created which could solve algebraic word problems and prove geometric theorems. These algorithms all generally followed the approach of stepwise heuristics to solve their problems and as such were saddled by the same limitation, namely traversing the enormous breadth of potential search paths in an efficient and timely manner (Russel & Norvig, 2003). A renaissance of sorts began in the 1980's with the advent of expert systems which attempted to use knowledge representations to mimic the decision-making capacity of a human expert. However, the optimism and promises of the expert systems again failed to meet expectations and enthusiasm and financial support both waned (Leondes, 2001). However, by the mid-90's many of the goals originally sought by artificial intelligence and machine learning began to see fruition, with one of the most notable being the creation of Deep Blue, a chess playing computer developed by IBM. After losing a series in 1996 to world champion Garry Kasparov, refinements were made by IBM engineers and in May of 1997 Deep Blue defeated Kasparov in a six-game rematch 3 ½ - 2 ½, becoming the first instance of a computer to best a reigning world chess champion (Higgins, 2017).

Recent Advances and Applications to Alzheimer's Disease

From these early implementations, access to large repositories of data, faster and better computers, and continued refinement of algorithms have pushed the field of machine learning even further and the discipline has seen an incredible amount of progress over the last 25-30 years. The early promises of the perceptron models are now reaching fruition with the advent of deep learning methods and multi-layer neural networks (LeCun et al., 2015; Schmidhuber, 2015) and notable advances in image processing, natural language tasks, machine translation, sequence analysis and many other fields have all begun to be realized. This includes medical domains such as Alzheimer's disease with particular emphasis on the potential role of machine learning in predicting changes in cognition.

Large-scale datasets with hundreds of patients have been created and include not only traditional patient characteristics like demographics, neuropsychological metrics, and measures of clinical function, but are further supported by more novel biomarkers from blood and spinal fluid, genetic assays and, perhaps most notably, a variety of neuroimaging techniques which directly measure brain structure and pathologies *in vivo* including magnetic resonance imaging (MRI) and positron emission tomography (PET) (Chen et al., 2014; Marti-Juan et al., 2020; Rathore et al., 2017). Furthermore, the progressive nature of AD has allowed these datasets to consist of longitudinal data with an eye towards the ultimate goal of slowing disease progression if not outright preventing disease transition. However, in spite of these datasets well suited for machine learning applications in AD, much statistical learning research has largely been limited to either cross-sectional designs or has only considered repeated measures with respect to covariates and not to outcomes (Chen & Bowman, 2011; Luts et al., 2012). This has largely been an issue with available methods and only recently has there been greater emphasis on fully leveraging the longitudinal nature of repeated measure patient data in machine learning paradigms. This is considered especially important in Alzheimer's disease as prediction of future cognitive states based on a patient's current clinical profile is of critical concern.

Supervised and Discriminative Learning

Regularized Regression

Of the many different types of machine learning paradigms, supervised learning methods have seen the most amount of research and development, most likely due to their reliance on well-defined outputs which greatly facilitates and directs the machine's task. Some of the earliest machine learning designs were simply adaptations of traditional, inference-based regression models which placed greater emphasis on increased accuracy of model predictions. These did so by leveraging the bias-variance trade-off in a model by reducing model variance at the cost of increased bias of coefficient estimates. The earliest of these methods were seen in 1970 with the proposal of ridge regression (Hoerl & Kennard, 1970), a method of constraining the coefficient estimates of highly correlated or non-important independent variables and shrinking them towards zero using a square ℓ2 penalty term. Subsequent adaptations include least absolute shrinkage and selection operator (LASSO) regression which is functionally similar to ridge regression but uses an ℓ1 penalty term based on the absolute value which allows coefficients estimates to actually reach zero for unimportant independent variables (Tibshirani, 1996). These two regularization techniques can also be combined using elastic net regularization which includes both ℓ1 and ℓ2 penalty terms with varying representative strengths which can be tuned specifically to the task at hand (Zou & Hastie, 2005). These methods are still widely used in a variety of disciplines in both academic and non-academic settings although most of their utility is taken with cross-sectional data on independent observations. However, extensions have been investigated to see how well these methods can generalize to correlated data structures such as longitudinal datasets. Methods to adapt both ridge and LASSO penalizations since they were first introduced have been attempted with the most common consideration being the inclusion of additional correlated structures such as the random effects components of linear mixed models (Skolov et al., 2016). Expanding on these refinements, applications directly related to medical research have been conducted on datasets such as those in longitudinal genome-wide association studies (Barber et al., 2017) and repeated measures in retinal optical coherence tomography data (Lang et al., 2016) as well as numerous simulation datasets. The combination regularization method of elastic net regression has also been investigated in a variety of medical research fields and is of great interest in reducing the domain space of biomarkers, for example in cardiovascular incidence (Eliot et al., 2011).

Despite these advances even within the general domain of medical research, applications of longitudinal regularization methods specific to neurodegenerative disease or cognitive decline have been sparse. Instead, most regression techniques which manipulate the bias-variance trade-off attempt to take some sort of adjustment to prior models that are simply informed or inspired by shrinkage and penalization. A recent example was dubbed likelihood contrasts where data were iteratively added to a standard mixed-effects regression model and the change in log-likelihood was evaluated according to classification to one of two groups. These models would then predict the classification based on the maximization, essentially using the log-likelihood as their objective function (Klen et al., 2020). Another recent design applied to psychology and inspired by machine learning methods is called Gaussian process panel modelling which is a more Bayesian approach to dealing with longitudinal data (Karch et al., 2020). While these implementations were evaluated using standard accuracy metrics, in both cases the predominant motivating rationale was to aid feature selection and not necessarily conduct prediction models based on the training data. Because of this variety in implementation and the relative dearth in neurodegeneration related applications, this has left characterization of these types of mixed-effects shrinkage methods in AD ripe for investigation. In spite of this potential, practical implementations of mixed-effects regularized regression have largely fallen out of favor in recent years, with libraries such as `GELMMnet` for Python and R packages such as `lmmlasso` and `lmmen` either being orphaned or no longer receiving regular update support as researchers have moved to other, more novel types of machine learning (ML) methods.

Ensemble Methods

Alternatively, ensemble methods, especially those explicitly using classification and regression trees, have seen much more research and support not only in their longitudinal extensions but also within the realm of AD specific research when compared to standard penalized regressions. Classification and regression trees (CART) are among the oldest classes of heuristic learners, first coined in 1984 by Leo Breiman (Breiman et al., 1984) to describe a directed acyclic graph which can be used to aid decision making. These trees provide the foundation for the ensemble methods wherein several trees are used in conjunction to create a "forest" to further increase predictive capacity of the models. These ensemble methods can be exceedingly varied with an early example being bagging, a portmanteau of bootstrap and aggregation, which builds multiple CARTs using several bootstrap samplings of the training data with replacement followed by aggregation across the forest of trees (Breiman, 1996). Another variation is boosting, with one of the most prominent examples being the Adaptive Boosting or AdaBoost algorithm, where trees are incrementally adjusted based on previous training instances to improve performance (Freund & Schapire, 1996). Another variation is the random forest which extends bagging from sampling subjects to instead sample various portions of the feature space and develop sets of independent variables from the data which reduces correlation within the forest and helps limit the tendency of decision trees to overfit to their training data (Breiman, 2001).

Ensemble methods are among the most popular of the supervised machine learning methodologies and as such have seen several attempts to extend into the analysis of longitudinal data. One of the earliest attempts was conducted by Hajjem et al. in 2011 (Hajjem et al., 2011) when they adapted regression trees to account for clustered data, including unbalanced designs. This was largely just an adapted expectation maximization algorithm that fit a regression tree instead of a fixed-effects parameterization but of greater interest was when the same group at HEC Montreal adapted the design further to create mixed-effects random forests instead of just single trees (Hajjem et al., 2014). Similar extensions have also been carried out with boosting methods for multivariate trees (Miller et al., 2017; Pande et al., 2017) as an alternative to bootstrap based methods like bagging and random forest while another method simply fused multiple mixed-effect trees together with one tree focused on fixed-effects and another random effects (Ngufor et al., 2019). However, in all of the cited articles there was no attempt to do characterization of these methods in the field of neurology, let alone in neurodegenerative disease and cognitive decline. For example, the motivating example in the mixed-effects random forest introduction was box office sales while the random forest fusions were focused on predicting hemoglobin A1c levels. However, despite these contextual limitations, longitudinal applications of ensemble methods are among the most popular paradigms with implementations in several programming languages including R (`randomForest`, `longituRF`) and Python (`merf`). These implementations have seen continual updates and improvements and are still considered ripe for further study unlike the relative quiet given to regularized mixed-effects models. As such, as these methods see continued support and investment by the machine learning field alongside greater traction specifically in medical research, they are a ready avenue of investigation for application in predicting outcomes in AD and neurodegeneration.

Kernel Methods

Among the most robust types of statistical learning models are those which make use of kernel methods, most notably support vector machines (SVM). Like many of the previously discussed methodologies, the SVM algorithm was first proposed long before it was refined and implemented in its current form. SVMs are rooted in the basics of Vapnik-Chervonenkis theory (VC theory) which was first proposed in 1974 (Vapnik & Chervonenkis, 1974) and provides a framework for machine learning explicitly from a statistical point of view with emphasis on consistency, complexity and control of generalization. SVMs are the most well-known practical implementation of VC theory and were first proposed as non-probabilistic linear classifiers in 1995 (Cortes & Vapnik, 1995). At the most fundamental level, they function similarly to the classic single-layer perceptron by separating and parsing the parameter space of a dataset. However, they extend on the perceptron by maximizing the distance separating the categories of a dataset in a classification task or the decision boundary that encapsulates the most data in a regression problem. In addition, SVMs are uniquely suited for non-linearly separable problems via the “kernel trick” which creates a non-linear hyperplane that maps the original dataset, which can consist of a high-dimensional feature space, to a reduced space that then uses the kernel to linearly parse the data implicitly (Bishop, 2006). A variety of kernels can be used with SVMs from rudimentary linear kernels to polynomial kernels with predetermined powers of exponentiation to radial basis function kernels whose expansion has an infinite number of dimensions. The particular strength of the kernel trick is that the higher dimension feature space does not need to be directly calculated and instead only the inner product space is necessary making calculation of the kernel function efficient and computationally tractable.

Due to their robustness and heavy use in the machine learning literature, support vector machines for classification and regression have also seen adaptations with respect to applications to longitudinal datasets. As with many of the early implementations, there was an initial limitation wherein response outcomes were only measured at single time points in a cross-sectional fashion rather than repeatedly or with other some sort of explicit structure (Chen & Bowman, 2011; Du et al., 2015). However, more recent advances have applied adjustments which can now account for true correlated or panel data including longitudinal responses. The common approach of these methods has been to expand upon the single kernel fitting inherent to SVMs and apply multiple kernels in tandem. These designs begin with one kernel, often with a standard implementation such as a Gaussian radial basis function, which focuses on solely modelling the feature parameterizations while another kernel, frequently structured to resemble a covariance matrix such as those seen in the random effects component of linear mixed models, instead handles subject-specific effects to account for the inherent correlation within the data measures (Chen et al., 2015). These multiple kernels are then fused together into a single linear function which can then predict either categorical classifications or real-valued outcomes in regression. These fusion kernels have an additional benefit of being to apply several separate kernels to the population-level effects as well, thus being able to have different kernels for different data modalities which may have disparate underlying distributions such as distributions of patient genetics versus population characteristics and demographics. Of interest, the mixed-effects SVM paper by Chen et al. (Chen et al., 2015) was applied to neurodegeneration and actually carried out on ADNI imaging data, specifically the MRI images, in order to predict a measure they referred to as brain age. In their more provincial forms, SVMs are very popular and widely used in a variety of disciplines when applied to cross-sectional data, in no small part due to being so intrinsically rooted in the fundamentals of statistical learning as a natural outgrowth of VC theory. However, this mathematical underpinning also makes them especially complex and less accessible when compared to other machine learning methods like regularized regression and ensemble forests. As such, practical implementations are rather limited with multiple kernel packages and libraries being especially sparse. Kernels for patient or cluster-specific effects are generally left to the user to design and implement as opposed to the commonly used and predefined radial basis function and polynomial kernels. Even supported packages such as `MKLpy` for Python and `RMKL` for R require significant amounts of user-provided definitions making generalized implementation especially challenging without extensive prior mathematical and statistical expertise. This is in turn reflected in the literature with very few published implementations of multiple kernel models on correlated data and with little advancement seen in the last several years as investigators have instead focused on other methodologies, specifically deep learning applications using artificial neural networks.

Neural Networks and Deep Learning

Many of the more recent advances in the field of machine learning have been developments in artificial neural networks (ANN) and deep learning. As a concept, ANNs build upon the original mathematical proposal of Pitts and McCulloch from 1943 (McCulloch & Pitts, 1943) to generate learning models inspired by the connective networks found in the biological brain. The nodes of an ANN represent the neurons while their connections mimic synapses and the strength or weights of these connections are adjusted by optimization of a loss function during the training process to either increase or decrease node connectivity as the system learns. Deep learning is an extension of ANNs using multiple layers of nodes to allow the system to extract higher level features using greater degrees of abstraction and representative knowledge at each subsequent layer and in turn learn more complex designs. In the decades since Rosenblatt's perceptron was implemented as a single-layer neural network (NN), several adaptations have led to the development of ever more powerful NN models to the point where deep learning is considered the current frontier of machine learning and artificial intelligence. Some of these early proposed improvements include the mathematical presentation of backpropagation in calculating the gradient of the loss function to efficiently train the connective weights of a network (Kelley, 1960), a method that is still used in feed-forward networks today, and the application of polynomial activation functions which allowed neural networks to be more than simple linear classifiers (Ivakhnenko, 1968). The 1968 paper by Alexy Ivakhnenko is especially pertinent as it also presented a multi-layer extension of the basic perceptron to enable the non-linear classification and is marked as the first instance of a deep learning system.

The early 1980's saw two especially critical developments in ANNs specifically in the context of longitudinal data representations. In 1980, Kunihiko Fukushima proposed the Neocognitron (Fukushima, 1980), the first convolutional neural network (CNN) which uses convolutional kernels to slide along the data inputs and create smaller, locally connected feature maps. Prior to the advent of CNNs, multi-layer networks had generally been fully connected with the nodes of each layer connecting to every node in the adjoining layers which often led to overfitting to the training data. By applying the convolutional kernels which emphasize proximal associations and local connections, CNNs can regularize their networks to assemble increasingly complex feature maps without overfitting. Another key adaptation was the development of the recurrent neural network (RNN), with the first implementations seen with the Hopfield Network in 1982 (Hopfield, 1982) and, its extension, the Boltzmann Machine in 1985 (Ackley et al., 1985). Prior to this, neural networks were unidirectional with a fixed depth, wherein a signal would only propagate in a single direction from one layer to the next. RNNs generalize this prototypical neural network structure by allowing nodes to self-connect, using an internal memory state to adjust these connection weights and simulate sequential or temporal behavior. RNNs are especially powerful as they can be either finite, with a limited number of recurrent edges which can be unrolled into a feed-forward network, or infinite with no limit on the number of internal connections. Both developments are especially key in the domain of longitudinal machine learning as they clearly demonstrate natural extensions to time-series data, either with the sliding convolutions mimicking time-step transitions or the recurrent neural network's internal updates as time steps. While several other adaptations have pushed deep learning methods as a whole, such as the resolution of the vanishing gradient problem (Hochreiter, 1998) and the application of graphical processors to speed up training through parallelization (Raina et al., 2009), these two extensions in particular were critical for longitudinal applications of ANNs.

Long Short-Term Memory Recurrent Neural Networks

After their initial presentation, a key issue was identified with especially deep NNs like the recurrent neural networks: the gradient problem. Standard RNNs are still generally trained using backpropagation methods, with the gradient of the loss function used to update weights after each pass through the system. However, with their relatively extensive depth, even for finite impulse RNNs, the gradients, which must still be calculated to finite precision, either vanish as they tend towards zero or explode as they tend to infinity. This is a problem for any ANNs of sufficient depth and was first identified by Sepp Hochreiter in 1991 (Hochreiter, 1991). Several solutions have been proposed to address the gradient problem such as the rectified linear unit (ReLU) as an activation function (Glorot et al., 2011) but an adaptation specific for RNNs was the development of long short-term memory (LSTM) RNNs. First presented in 1997 (Hochreiter & Schmidhuber, 1997), LSTM RNNs address the gradient issue by allowing errors to propagate through the self-connecting edges unchanged. It does so through the use of “memory gates” which can direct how a recurrent unit can either update a weight based on the gradient, leave it unchanged, or outright exclude the error in a reset function. Even after their initial presentation several decades ago, the standard LSTM remains one of the best starting points for handling error propagation in networks using recurrent units and are still used today to help initialize more complex implementations of RNNs (Le et al., 2015). With many of the early issues with proper training resolved and their natural application to sequential datasets, it is straightforward to see how LSTM RNNs can easily generalize beyond their natural language processing origins and are especially well-suited for applications to longitudinal datasets.

One-Dimensional Convolutional Neural Networks

As mentioned, one of the earliest efforts to help regularize fully connected neural networks was with convolutional neural networks. Mathematically, convolutions are simply filters applied to a set of inputs which are then used to create a feature map of local areas which can overlap to represent the entire dataset in a more abstract fashion while minimizing information loss. The key principle of CNNs is that by applying these filters in a sequential fashion, the CNN can identify spatial and temporal dependencies within the input by putting greater associative strength on more proximal convolutions. This reduces the dimensional space of the original input, creating more and more abstract and computationally tractable feature maps, without losing the internal dependencies of the data (Goodfellow et al., 2016). After convolution, there is then a pooling layer which further reduces the dimensionality of the feature map making calculations even more feasible. This process is then repeated to develop an abstract representation of the input which can later be used either for reconstruction of the training data (e.g. for denoising of images) or generalization to feature extraction and identification in new data.

The original application of the CNN was the previously described Neocognitron, which was developed for image processing. In fact, processing and feature extraction of images are still the most widely used applications of CNNs, with each convolutional filter window stepping through the two-dimensional image to create the feature maps. However, researchers have also generalized the CNN to be applied in datasets outside of images, developing networks that still leverage the internal structure that is critical to capture, as desired in sequential and time-series data (Kiranyaz et al., 2021). The general idea remains the same, with shifting of a convolutional filter to extract proximal dependencies followed by pooling to reduce dimensionality and improve calculations, but the specific architecture and mathematical operations used in training emphasize associations in a single dimension. This has enabled extensions of CNNs beyond its image processing origins and let them be used in a variety of settings like human activity signal processing in wearable devices (Lee et al., 2017) and encouraging the use of one-dimensional CNNs (1D CNN) to other directionally structured datasets including longitudinal applications.

Deep Learning Methods and Medical Research

While there has been substantial progress in the field of self-supervised and deep learning, many of the advances have taken place outside the context of medical research (Liu et al., 2018; Liu et al., 2020). As discussed, much of the contextual focus has been placed on image processing, recognition, and feature extraction, especially for convolutional neural networks. Similarly, much of the work using LSTM recurrent neural networks has focused on natural language processing and machine translation, leveraging the sequential relationships inherent in language (Palangi et al., 2016). Although there have been extensions outside these more typical domains, such as the analysis of high dimensional time series data in biometric tracking (Ravi et al., 2016) or the previously cited example in wearable devices, it is only recently that adaptations of self-supervised and deep learning methods have been applied to medical research. Some of these extensions are very natural, such as the classification of temporal electrocardiogram signals in cardiology (Singstad & Tronstad, 2020) using CNNs or leveraging LSTM RNNs to identify epileptic seizures using electroencephalogram data (Xu et al., 2020). However, these applications tend to resemble the original applications, with a high density of small sequential increments which mirror the proximal pixel relationships of images or the smaller steps of continuously recorded time series inherent to human activity data. As such, the ability of these deep learning techniques to generalize to domains like clinical trial data with more disparate and disjoint time courses, like those observed in studies of neurodegeneration, has been rather limited and it is largely unknown how these self-supervised methods will work with AD data. Regardless, there is a great deal of potential to utilize these sequence dependent methods which may garner additional predictive capacity beyond what is possible using standard supervised machine learning methods.

Non-Longitudinal Machine Learning in Alzheimer's Disease

As has been shown, much of the emphasis on longitudinal aspects of statistical and machine learning has had a focus outside of medical and clinical applications, let alone within the domain of neurodegeneration and Alzheimer's disease. However, that is not to say that machine learning paradigms have not been widely applied to AD. Machine learning designs to aid in diagnosing dementia were seen as early as 2008 when SVMs were used to distinguish AD from normal aging and fronto-temporal dementia with the results compared to diagnoses by human radiologists (Klöppel et al., 2008). This has not wanted as a simple search in PubMed for "machine learning" and "Alzheimer's disease" returned 330 publications in 2020 alone. Since those beginnings, an enormous number of studies have attempted to leverage statistical learning to either improve diagnosis or to create metrics that can be used as aggregates or surrogates. Some studies have attempted to use dimensional reduction and feature selection for idiopathic AD, with some studies attempting to find genetic variants beyond the well-known *APOE4* allele (De Velasco Oriol et al., 2019; Huang et al., 2018) while others focus on identifying novel proteomic biomarkers (Bader et al., 2020). Taking inspiration from the gut-brain axis implicated in Parkinson's disease, studies have even begun applying high dimensionality techniques to the gut microbiome in AD (Kaur et al., 2021). In addition to feature selection, prediction and diagnosis using machine learning methods have also been evaluated in AD. The potential of machine learning when applied to neuroimaging has long been recognized and investigation into imaging-based classification has been encouraged for some time (Mirzaei et al., 2016; Rathore et al., 2017). Furthermore, with such a wide variety of cross-sectional methods, some studies have taken comprehensive approaches to compare different machine learning methods to predict either diagnosis state or age of onset (Naik et al., 2020) or review different novel learning methods which have been previously considered (Marti-Juan et al., 2020). Some studies have even attempted forecasting to predict slopes of change of neuropsychological metrics as markers for AD progression (Fisher et al., 2019). However, these studies all generally suffer from the same sets of limitations. The vast majority focus solely on cross-sectional measures using independent data units with no consideration of longitudinal structure. While this is valuable for identifying features or biomarkers associated with categorizations of neurodegeneration and AD, they do little to inform prediction within units. Even among those that do make use of longitudinal data, many rely on aggregate measures, often calculating metrics like annualized change, which are then used as the outputs using the same types of cross-sectional machine learning methods requiring independent inputs. Attempts to utilize actual longitudinal panels and repeated measures of patient data are exceedingly rare which limits the ability to use generative models to predict disease progression, cognitive decline, and functional transition. Critically, this not only holds for novel observations and patients with previously unseen characteristics, but also restricts the future prediction and forecasting of metrics and AD classifications within patients based on their prior observation profiles. Although AD research has seen enormous strides from the advances in machine and statistical learning, this lack of characterization using true longitudinal models which fully leverage the structure and correlation of longitudinal data is a major deficiency in the literature and a deeper, more comprehensive investigation is a feature the AD research field has been eager to receive.

Special Considerations for Subject-Specific Effects

The desire for longitudinal prediction of ADRD outcomes is abundantly clear and, as has been discussed, models which are able to utilize these data structures are central in furthering research in the field. Even with the more standard inferential methods, there is an appreciation that care is needed when modeling these data, as rudimentary statistical designs are unable to adequately account for within-panel relationships. For example, basic ordinary least-squares regression assumes all observations are independent and ignores the interrelatedness of repeated measures, leading to potential errors in inference and incorrect conclusions (Bernal-Rusiel et al., 2013; Burton et al., 1998; Fitzmaurice et al., 2011). Internal similarity is expected with repeated measures, referred to as serial correlation, and modelling of these subject-specific effects is critical for proper analysis of longitudinal data (Higgins et al., 2001). Fortunately, several designs allow for panel data and are leveraged in ADRD research. Prevalent is mixed-effects modelling using the population-level fixed effects seen in common regression techniques alongside subject-specific random effects components allowing both starting points (intercepts) and trajectories (slopes) to vary on an individual-by-individual basis (Donohue & Aisen, 2012; Doody et al., 2001). Mixed effect models are invaluable in ADRD research, aiding inferential conclusions in both interventional trials and observational studies (Ard et al., 2015; Gavidia-Bovadilla et al., 2017).

With refinement of statistical techniques, method evaluations have used a combination of long-running natural history datasets and clinical trials from the literature (Capuano et al., 2018; Ito et al., 2013; Rogers et al., 2012) alongside data simulation studies to assess model generalizability across more varied scenarios (Chen et al., 2018; Di et al., 2016; Wang et al., 2018). Many have shown potential with prediction or generation of trajectories of decline in ADRD (Kim et al., 2021; O'Shea et al., 2021) but some aspects remain under-investigated, such as the role of participant-level effects and their impact on estimation of cognitive outcome measures (Li et al., 2020). A critical point is exact subject-specific effects are unique to an individual within a given model and must otherwise be imputed based on model covariances, a practice which may not be appropriate (Giil et al., 2021; Guo et al., 2021). Furthermore, models are developed with focused and specific goals but when leveraged outside their original purpose may have unforeseen consequences. For example, pre-specified parameterizations may behave well under certain assumptions but be severely misrepresentative in scenarios for which they were never intended (Breitve et al., 2018; Giil & Aarsland, 2020; Milliken & Edland, 2000; Uspenskaya-Cadoz et al., 2019). Understanding the influence of subject-specific effects under certain designs, such as imputation, can assist subsequent model creation and direct how to best use calculated predictions of cognitive outcomes.

Deeper and more refined knowledge of the performance and behavior of these different model designs, types of predictions and forecasts, class of outcomes, and various subject-specific effects can all tie together to help direct researchers to make the most informed decisions possible when developing their own research methods. It would be entirely expected for each combination of model design to have their own set of strengths and weaknesses and the preferred model is dependent on the goals of the investigator. Evaluation of the influence of subject-specific effects is a critical step in the process of characterizing predictive performance of longitudinal machine learning methods in Alzheimer's disease, providing both a current evaluation of the field and a pipeline for future machine and statistical learning paradigms.

METHODS AND APPROACH

Outcomes

Two outcome measures were used for evaluation of the machine learning methods in this dissertation to provide a more comprehensive assessment of their predictive capacity, one continuous and one binary categorical. The continuous measure for regression analysis was the Alzheimer's Disease Assessment Scale – Cognitive Subscale (ADAS-Cog) while the categorical outcome was a binary recasting of the Clinical Dementia Rating (CDR). Two responses were utilized to not only evaluate the ML methods on multiple classes of variables but also provide two different outcome contexts in ADRD research.

The ADAS-Cog, the continuous outcome for regression analysis, is an assessment battery designed to evaluate multiple domains of cognitive functioning including memory, reasoning, orientation, praxis, language, and word finding difficulty. It is scored on a continuous scale ranging from 0 to 70 total errors with higher scores indicative of greater levels of impairment. It is commonly used as an outcome in AD clinical trial settings as a more refined assessment of both global and domain specific cognition. Like other neuropsychological assessments such as the Montreal Cognitive Assessment (MoCA) and Mini-Mental State Exam (MMSE), the ADAS-Cog has higher resolution at greater levels of impairment and demonstrates better ability to distinguish severity among impaired individuals. However, its sensitivity is compromised when attempting to assess the cognitively intact, losing this resolution at lower scores. In addition, the ADAS-Cog is largely a tool for research purposes and due to its length and involved administration is generally not used in clinical settings. However, research has indicated a 3 to 4 point increase in ADAS-Cog over a six-month period can viewed as clinically meaningful change, at least in early AD.

The CDR is a five-point scale which characterizes six domains of both cognitive and functional performance including memory, orientation, judgment, community affairs, home & hobbies, and personal care. Because it assesses both cognitive and functional ability, it can be used as a diagnostic tool in clinical settings or as a metric in ADRD studies. Specifically, it distinguishes between a cognitively intact individual (CDR score of 0), an individual with mild or questionable impairment (CDR score of 0.5), or someone who has advanced beyond cognitive difficulties to varying degrees of the functional impairment observed in dementia (mild, moderate, and severe dementia with a CDR score of 1, 2, or 3 respectively). In particular, the CDR is able to draw diagnostic conclusions on mild cognitive impairment in the absence of functional deficiencies since perturbations observed only in the memory domain results in scores of 0.5 indicating mild impairment. For the purposes of this study, in order to provide a counterpoint to the later stage resolution of the ADAS-Cog, CDR scores were binned at a breakpoint of 0.5 to create a binary outcome (non-impaired vs impaired) and specifically emphasize earlier stages of cognitive decline.

The use of these two outcomes imparts several key benefits. First, it allows for an assessment of the machine learning models when they are utilized for both regression and classification purposes, both of which are common within ADRD research as well as for statistical modelling in general. They also provide multiple contexts for Alzheimer's disease as the ADAS-Cog subscale and classification by CDR have their own utility in dementia research. The ADAS-Cog is primarily a research tool while the CDR is designed for diagnostic ability and is much more likely in clinical settings. As previously discussed, there are also temporal contexts for how each response was used in this study wherein the ADAS-Cog provides high resolution on distinguishing levels of impairment among those who are exhibiting some level of dementia already while the CDR is much more capable at identifying more mild levels of impairment, including distinctions between the cognitively intact and those with only mild impairment. This leads to an especially important characteristic of these two outcomes: clinical vs research utility. The ADAS-Cog as a metric is much better suited for researchers in ADRD but the scale itself may have relatively little importance in a clinical setting as the notion of a "one point change" in ADAS-Cog or score shifts below the cited 3-4 point change may not especially meaningful to a patient. Conversely, diagnostic categorization using the CDR may be viable as a study inclusion tool but may be too coarse of an outcome when conducting an interventional clinical trial focusing on more subtle changes. However, being able to identify a patient as impaired versus non-impaired may have much more clinical utility for both patient and provider, especially if this classification can be predicted at future timepoints for an unimpaired individual.

Reference Models

In addition to cross-model evaluation of the ML implementations, all models were initially contrasted against more standard inferential methods as control models using pre-specified parameterizations. This was to establish baseline comparisons with the specific hypothesis that the ML models would present with superior predictive capacity by virtue of having improved outcome metrics relative to the reference designs. For both the ADAS-Cog and CDR impairment outcomes, parameterizations were established outside the evaluation dataset used for the ML training and testing.

The reference model for the ADAS-Cog was taken from the parameterization of the Critical Paths for Alzheimer's Disease (CPath) consortium first presented by Rogers et al. in 2012 (Rogers et al., 2012). The CPath model was developed from a variety of literature reported values and cohort studies to describe progression of the ADAS-Cog in both natural history and randomized clinical trial settings with the goal of creating a framework to generate representative simulation cohorts which could be used for feasibility purposes when designing future interventional or observational studies. The model uses the parameterizations from a beta regression mixed-effects design to accommodate the bounded ADAS-Cog score by transforming the natural 0-70 score range to a 0-1 normalized scale. Initial ADAS-Cog scores are created using baseline MMSE score with longitudinal trajectories according to baseline age, sex, *APOE4* allele count, and baseline MMSE. Additionally, subject-specific effects can be randomly sampled using the provided model covariance matrices for both intercept and slope. Model parameters were developed using both summary-level and patient-level data using a Bayesian implementation to adjust meta-data from the literature with individual-level effects. Although the original model was tuned out to two years of linear time, it has demonstrated effectiveness when used for wider times frames. Further details about the CPath model can be found in Rogers et al. (Rogers et al., 2012) as well as an implementation in R using the `adsim` package (Polhamus, 2013), including coefficient values for population-level covariate effects along with covariance measures used to generate subject-specific effects.

Unlike the CPath model for ADAS-Cog, a pre-specified parameterization for the categorical casting of CDR was not available from the literature. Instead, a similar mixed-effects regression model was developed using a holdout subset of the multi-study meta-database used in this dissertation. After harmonization of the dataset, as described below, 15% of the CDR measures were extracted with subjects representatively sampled such that their final timepoint of evaluation was reflective of the terminal times for the entire dataset. These CDR values were utilized solely in the development of the reference parameterization and were never used during the training, tuning, or testing of the ML models. Much like the CPath reference model, the CDR impairment reference used a mixed-effects model, although specifically using a logistic regression design to accommodate the binary outcome. The fixed effect covariates used in model building were the same as those used by the CPath model, specifically baseline age, sex, *APOE4* allele count, and baseline MMSE with a linear time component. Unlike the CPath model there was no time-MMSE interaction term as this was found to impede model convergence due to overfitting of the model. An unstructured random effects design was used with subject-specific effects for intercept, slope, and a covariance term between the components. CDR impairment modelling was done using the `lme4` package in R (Bates et al., 2015).

For both reference model designs, only the population-level coefficient parameters were used to calculate predicted responses. Although values for the model covariances were available for each reference, these were not used when predicting outcomes, except for the imputation design for the ADAS-Cog in aim 3 as described below. In addition, fitted subject-specific effects were never used during forecasting as these were not available given the pre-specified nature of the parameterizations.

The Alzheimer's Disease Meta-Database

Data for this dissertation was drawn from a meta-database consisting of 18 clinical trials from the Alzheimer's Disease Cooperative Study (ADCS) and the four phases of the Alzheimer's Disease Neuroimaging Initiative (ADNI) observational study. The meta-database was first presented in 2014 by Kennedy et al. (Kennedy et al., 2014) and has participants representing the full spectrum of Alzheimer's dementia, from cognitively intact to diagnosed clinical AD. In its base form, the meta-database consists of 8936 participants with nearly 46,987 observational timepoints extending out to 12 years and includes study subjects with longitudinal data as well as single timepoint participants who were screened for inclusion but were not involved with the final studies.

Harmonization of the dataset was required in advance and was designed to map disparately coded visits to a continuous temporal variable based on study date, including 2392 participants with only baseline data or screening-only subjects. Key aspects of the data preparation include the requirement of the principle ADAS-Cog and CDR outcomes which were not available in all studies. Furthermore, all subjects were required to have the population-level demographics used in the reference models, most notably genotyping for *APOE4* allele counts. Although additional timepoints were available, the follow-up time was capped at six years as the ADCS clinical trials were never conducted beyond three years and the longer-term ADNI participants were found to be inherently unimpaired and were anticipated to impact generalization of model prediction at more distal timepoints. Full details of the data preparation and harmonization process can be found in the flow diagram in Figure 1.

The final covariate feature sets used in the ML models of aims 1 and 2 included baseline covariates of baseline age, sex (male, female), race (3 level factor: White, Black/African American, Other), ethnicity (2 level factor: Non-Hispanic or unknown, Hispanic), education (6 level factor: did not graduate high school (HS), high school diploma or general equivalency degree (GED), some college, college degree, some post-graduate education, post-graduate degree), any use of anti-dementia medication (donepezil, rivastigmine, galantamine, tacrine, or memantine), and *APOE4* allele count. Time dependent covariates included MMSE score, weight, and systolic and diastolic blood pressure with missing covariates imputed using last observation carried forward as needed.

Types of Prediction and Establishment of Training and Testing Sets

For all aims, two types of predictions were considered: development of whole-subject temporal profile trajectories and forecasts of final observations given prior data. The whole-subject profiles, colloquially referred to as trajectories, predicted outcomes at all observed timepoints for a given subject, from baseline to their final visits. Forecast predictions of final observations, also referred to simply as forecasts or forecasting, only considered the final observation for an individual and expressly made use of prior sequences of outcome responses where applicable, such as direct use of fitted subject-specific random effects for individuals used in model generation.

To accommodate both cases of predictions, two types of holdout datasets were created for testing of the models. After harmonization of the meta-database, including the previously described 15% extraction of CDR measures to create the classification reference parameterization, the meta-database had 10% of the remaining 3501 subjects randomly selected as the holdout testing set for trajectory evaluation, with all timepoints extracted. Of the remaining 3118 subjects, 10% of those were then randomly selected for forecast observation evaluation, with only their final timepoint extracted. For forecasting assessments, the preceding data for these subjects was used directly, whether as fitted subject-specific random effects for ML models which used mixed-effects regression components, or directly leveraging the previously observed sequences for the deep learning methods. The same testing set was used for all model evaluations with the remaining data used solely for training purposes in aims 1 and 2. Full details of the establishment of the holdout testing sets can be found in the flow diagram of Figure 2. Details of the baseline covariates described in the previous section for the final training and testing sets can be found in Tables 1 and 2 for continuous and categorical variables respectively.

Cross-Validation for Hyperparameter Tuning

The various ML models used in this study also contained several hyperparameters, configurations external to the model in question whose values cannot be estimated from the data and are instead heuristically determined. Hyperparameters serve several critical roles. One is improvement in model performance leading to more accurate predictions. Another is improvements in model efficiency to make computationally challenging models, due to lengthy processing times or extensively large memory footprints, more tractable. Hyperparameter tuning is a critical step in machine learning and was a necessary component of proper model evaluation for this study’s aims. Hyperparameter tuning and optimization followed a standard grid search protocol with hyperparameters for each ML model specified in advance and each combination of parameters evaluated in turn. Evaluation of each hyperparameter set used 10-fold cross validation, with the exception of the boosted trees ensemble method which used 5-fold cross-validation for computational tractability. After establishment of the training dataset described above, subjects were split into 10 equally sized groups or cross-folds. Each of the 10 folds was then sequentially held out during model generation and used as an internal validation set to calculate predictions for data not used in the model generation, yielding 10 distinct models with 10 sets of testing predictions and evaluation metrics which were then averaged for a final performance assessment. This was repeated for each of the hyperparameter sets with the final model building set either selected by consensus of metrics with the best values or manually selected when consensus was unclear. In those later cases, selection generally prioritized minimal root mean square error or mean absolute error for ADAS-Cog regression or maximal area under the receiver operator characteristic curve (ROC AUC) or recall sensitivity for CDR impairment classification. The 10 models from that hyperparameter set were then used for evaluation of the holdout testing datasets with the predicted ADAS-Cog values or impairment classifications scores averaged across the models as per the recommended procedure by Hastie et al. (Hastie et al., 2009). The same collection of cross-folds was used for all models, with the exception of the boosted models which combined adjacent groups (i.e. first fold with second, third with fourth, etc.) for its 5-fold design. Different hyperparameters sets were considered for regression and classification models independently; however, hyperparameter selection was only conducted on trajectories, with the same tunings used for both whole-subject profiles and final observation forecasting. Finally, the hyperparameter set identified for the ADAS-Cog for the mixed-effects random forest in aim 1 was the same set used during evaluation of subject-specific effects assessments of aim 3. Individual hyperparameters for each ML method are described in their corresponding section below.

Evaluation of Model Predictions

To quantify the performance of the evaluated ML models and the inferential reference standards, the model predictions for the ADAS-Cog and CDR impairment classification were compared against the known values from either the cross-fold holdout during hyperparameter tuning or the relevant holdout testing set during final model evaluation.

The regression metrics included root mean square error (RMSE), the square root of the mean squared difference between the predicted and true values, and the mean absolute error (MAE or mean AE), the mean of the absolute distance between the predicted and true responses. These metrics comprise both the variance and the bias inherent in a predicted outcome with both being utilized wherein the RMSE is more common, but the MAE is less prone to influence by outlier values due to extreme values of either predictions or true outcomes. The symmetric mean absolute percentage error (SMAE%), the absolute difference between predicted and true values divided by half the sum of those values, was also calculated but was only used during the hyperparameter consensus process as a potential tie-breaker metric. The bias was also calculated in two fashions: the raw bias, taken as the difference between the predicted and true values which can be negative, and the absolute value of the bias (AVB or AV bias) which forces the bias to be positive. These metrics assesses deviation from the expected value of an estimator and comprises the non-systemic component of model prediction error with the bias helping determine over or underestimation while the AVB is a more robust statistic for evaluation. Importantly, both errors and biases are scale dependent with smaller RMSE, MAE and AVB all indicative of superior model performance while raw biases closer to zero are preferred regardless of sign.

The classification metrics were largely based on the layout of each model's confusion matrix which tabulates the number of true and false positives and true and false negatives with nearly all metrics comprising some combination of these counts. The first metric was accuracy, which is the proportion of correct predictions (the sum of true positives and true negatives) out of all predictions. It is known that accuracy can be an inappropriate evaluation metric in isolation and can be highly misrepresentative in imbalanced datasets. However, since all reference and ML model evaluations involved relative comparison of performance metrics, this was considered only a mild concern. The other confusion matrix metrics included precision, also known as the positive predictive value, which is the proportion of true positives out of all predicted positives (i.e. true positives divided by the sum of true positives and false positives) as well as recall, better known as sensitivity, which is calculated as the proportion of correctly identified positives out of all predicted positives (i.e. true positives divided by the sum of true positives and false negatives). These two metrics are less prone to imbalance than accuracy and answer two distinct questions about a classifier's performance. Precision considers how many of the selected responses are relevant and is a critique of a model's tendency towards false positive selection while recall asks how many relevant responses are selected and provides an assessment of a model's tendency to misclassify false negatives. All three metrics are calculated as normalized proportions with higher values indicative of improved prediction performance.

Importantly, all three metrics are dependent on selecting a cut point to categorize the outcome response score of a classifier model to a binary class label. Optimal cut points were calculated in a data-driven fashion on an individual model basis. Specifically, densities of the impairment score classifiers were calculated, and the two largest peaks identified (i.e. the two values around which the response scores tended to cluster). The midpoint between these peaks was then used as the optimal cut point with scores below the cut point cast as non-impaired 0's while scores above the cut point were cast as impaired 1's. In some cases, the combination of classifier imbalance and large feature space made peak isolation, especially for non-impaired subjects, challenging. In those cases, Youden's J statistic was instead used as the optimal cut point. Youden's index considers the sum of the sensitivity and specificity at all possible cut points for the scores of a classifier with the optimal value being the score which maximizes this sum. Regardless, because of this dependency on a specified cut point for accuracy, precision and recall, receiver operator characteristics (ROC) were calculated for the classifiers using the standard metrics of sensitivity and 1 minus the specificity. Normalized areas under the curve (AUC) were calculated using the polygon rule and provided as an additional performance metric of global classification performance with the benefit of being cut point agnostic to a cut point.

A final metric was the net reclassification improvement or index (NRI) which quantifies how well a new model reclassifies an outcome, in terms of changing an incorrect response to a correct one and vice versa. The tabulation process of the NRI can be either positive or negative, with a negative value indicating worse overall reclassification performance of the new model, with higher values representing classification improvement. This metric was specifically used when comparing the various ML classification models against the logistic regression reference model and was not used during hyperparameter evaluation. There has been some concern in the literature about the use of NRI as a metric as even uninformative covariates may lead to positive NRI values much in the same way additional parameters improves coefficients of determination in a regression model (Pepe et al., 2015). However, preliminary investigation found that when considering the improvement of a classifier with inclusion of additional covariates to a model, there was no issue with using the NRI, so long as the assessed models were fit on a different dataset than the data being reclassified. However, when comparing models using the same dataset they were built on, such as for diagnostic purposes, other metrics such as the standardized net benefit are preferred as they are more apt to penalize the inclusion of non-associated variables. Although this could be a concern during features selection in the high-dimensionality feature spaces of machine learning models it was found to not be an issue for the current study as independent datasets were used for model training and hold-out testing thus making the NRI an adequate comparison metric for CDR impairment classification in this context.

Ensemble Machine Learning Methods

Although both regularized regression models and support vector machines with modified kernels were initially considered as prospective supervised methods for aim 1, adequate and well-defined implementations could not be secured. Packages and libraries were either underdeveloped and lacking in the necessary features for this dissertation work or were no longer supported by the authors as more advanced techniques and methods became preferred. Over the course of the study, both designs were eventually abandoned, and emphasis was instead given to the ensemble methods which continued to receive support from both developers and the ML research community.

The final set of evaluated machine learning models focused on ensemble tree-based methods. This included mixed-effects random forests (MERF) which sample from the feature space, bootstrap aggregated (bagged) generalized linear mixed model (GLMM) trees which sample from the set of meta-database subjects, alongside a single non-bagged GLMM tree for comparison, and sequential boosting of residuals (boosted) mixed-effects trees. These trees all follow the same general design where population-level effects are determined by the tree while subject-specific effects are modelled in the terminal leaf nodes. For all trees, the objective functions were optimized using the squared error loss for ADAS-Cog regression and cross-entropy loss for CDR impairment classification. In addition, hyperparameter tuning was done on a per-model basis in order to individually optimize each model according to its unique tuning profile.

Mixed-Effects Random Forests

Random forest models are ensemble methods which improve upon standard decision tree designs by allowing for "feature bagging" to randomly select a subset of model features and generate a forest of partial feature set trees which are uncorrelated. Tree outputs are then averaged across the forests to improve overall predictive accuracy and limit the need for tuning of tree-specific hyperparameters. MERF models extend the random forest by including mixed-effects models in terminal nodes to accommodate the serial correlation inherent in repeated measures and panel data. These additional components provide subject-specific effects which then update the population-level effects in the random forests stochastically. Hyperparameters include the proportion of the feature sets randomly selected to build each partial tree (set at 20%, 40%, or 60% of the feature space for ADAS-Cog and 35% or 70% for CDR-based impairment) and the number of trees built for each random forest (set at 250, 500, or 750 trees for ADAS-Cog and 500 or 750 trees for CDR-based impairment).

The current implementation for this work was based on the design presented by Capitaine et al. (Capitaine et al., 2021) in the `longituRF` package in R (Capitaine et al., 2021) with modifications to provide greater control over how the subject-specific effects were used during prediction. The original functionality only allowed for predictions using fitted random effects, requiring subjects to have been used during the model building process. Accordingly, the models could not accommodate prediction on new data. The predict function was rewritten with the additional modifications expanding on this basic use of known subject-specific effects, which was used for observational forecasting. These adaptations also allow for complete suppression of the subject-specific effects component to rely solely on the population-level fixed effects or to impute subject-specific effects based on parameterizations of the model covariance matrices as desired.

Bagged Generalized Linear Mixed Model Trees

Generalized linear mixed model trees are a modification of standard classification and regression trees which explicitly account for the clustered structure of panel data such as in repeated measure data (Fokkema et al., 2018). Similar to the tree building described for the MERF model, trees are first built based on the fixed effect parameterizations of the model before application of a random effects component in the terminal nodes which is then used to update the tree using recursive partitioning. Bootstrap aggregation, or bagging, is a generalized procedure for ML models which fits multiple models simultaneously after sampling with the collection of outputs averaged across all trees, limiting the requirement for tree pruning. Unlike the covariate subsets of MERF models, bagging procedures use the full feature space but instead randomly sample with replacement from the data points themselves, in this case, the meta-database subjects. Panels which are not selected (the out-of-bag set) are then used for model validation and tuning. Similar to the MERF design, the tree-specific tuning is less required and the considered hyperparameters included the proportion of samples used for each tree (set at 40% or 75%) and the number of generated trees (set at 100 or 200 trees).

The basic GLMM implementation for this study used the `glmertree` package in R (Fokkema et al., 2018) and was used for creation of the set of trees built during bagging as well as the single GLMM tree used as a reference. The process for conducting the bootstrap aggregation, including the random sampling of the dataset, averaging across the bagged GLMM trees, and validation of the results using the out-of-bag samples, was developed independently, and applied as a wrapper to the primary GLMM function.

Boosting on Mixed-Effects Trees

Unlike the sampling processes used by the MERF and bagged ensemble methods, boosting uses the entire feature space and sample set when building its trees. Instead, the output of a tree is used to modify the dataset and this new dataset is used to iteratively build the next set of trees. Boosted decision trees are the specific extension of Jerome Friedman's original gradient boosting machine (Friedman, 2001), whose most common implementation is the widely used AdaBoost algorithm, into a classification and regression tree architecture. Boosted mixed-effects trees further extend this architecture to allow for panel data by including mixed-effects regression models in terminal nodes prior to the boosting stage. Unlike the sampling based MERF and bagged GLMM models, outcomes are more dependent on tree structure and pruning via hyperparameter tuning is required. Additionally, as trees are fit iteratively to convergence, performance hyperparameter optimization and parallelization becomes critical as model building can be especially time consuming with exceptionally large memory demands. The hyperparameters include the number of trees, the minimum number of samples in the terminal nodes, the layer depth of the trees, and the shrinkage rate applied to the residuals during boosting.

Implementation of the boosted trees used a combination of the `gbm` and `mvtboost` packages in R (Hickey et al., 2016; Miller & McArtor, 2017). The boosted trees were found to be especially computationally intensive which limited their implementation and tuning in this dissertation. As previously described, hyperparameter tuning could only allow for 5-fold cross-validation. Tuning was also limited to considering the number of trees used in boosting and was set at either 100 or 200 trees. The other parameters, terminal node samples (20), tree depth (15), and shrinkage rate (0.01), were instead based on performance tuning conducted during pilot testing on a smaller subset of the data with a reduced feature space. These hyperparameters were found to perform adequately for both ADAS-Cog regression and CDR impairment classification and were used for all boosted models. Additionally, while cross-validation was able to be conducted in a regimented fashion for most of the ML methods in this study, the boosted tree models were instead run independently in parallel in order to accommodate these specific memory and time demands with additional wrapper scripts developed to assist in this process.

Deep Learning Neural Networks

The selection of artificial neural networks used in aim 2 included recurrent neural networks using a long short-term memory component for sequential data (referred to here as LSTM RNN or simply LSTM or RNN), a one-dimensional convolutional neural network for sequences (1D CNN or CNN), and a standard non-sequential feed-forward neural network (FNN) as a comparison control. All ANN models used standard build methods with connection weights between nodes calculated using stochastic gradient descent during back propagation. A rectified linear unit (ReLU) was used as the activation function between connecting layers while a sigmoid function was used to calculate the final output score which was then transformed to either an ADAS-Cog regression value or cast to an impairment classification. Like the ensemble methods of the first aim, the optimized objective function was mean square error loss for ADAS-Cog regression and cross-entropy loss for CDR impairment classification and 10-fold cross-fold validation was used for hyperparameter tuning. Of note, R does not have native neural network functionality and instead coordinates with other programming languages which build the models. This dissertation used an interface with Python 3.3 using the `reticulate` package (Ushey et al., 2021) while network building was done using TensorFlow via the `tensorflow` package (Allaire & Yuan, 2021) with Keras as the facilitating interface using the `keras` package (Allaire & Chollet, 2021).

Non-Sequential Feed-Forward Neural Networks

To compare against the sequential ANN methods, a standard multi-layer perceptron was created as a control design. Time was still included as a feature but was simply considered as another covariate with each timepoint observation considered an independent event. All network layers were densely connected with the number of nodes per layer decreasing by 50% at each subsequent layer. Hyperparameters included the total number of densely connected layers (set as 2 or 3) and the number of starting nodes in the first layer (set as either 8 or 4). Other hyperparameters were available for consideration including constraints on allowed values for the kernel connections and starting bias values. Initial testing indicated benefit of constraining the kernels but not the bias with the preferred constraint being a maximized normalization constraint with a maximum value of 2. These constraint settings were used for all node and layer combinations used during tuning. In addition to the full temporal FNN built for model evaluation, a baseline FNN was created using only values at baseline for both types of outcomes. This baseline FNN was used for prediction of trajectories in the sequential NN models as described below.

Long Short-Term Memory Recurrent Neural Networks

Recurrent neural networks are extensions of basic feed-forward networks with the inclusion of a cyclic internal state to retain information about variable sequences of inputs. Long short-term memory adaptations extend this further by incorporating a gating feature which allows the network to retain both distal and proximal information about prior sequences independently using “forget gates” which control the amount of sequential information that is retained both short-term and long-term. In addition to allowing both types of prior sequencing to be used during prediction, it resolves the vanishing gradient problem inherent in RNNs when applied to longer sequences. Although LSTM RNNs have their roots in natural language processing (Jozefowicz et al., 2016) their unidirectional behavior have natural extensions to temporal data as used in this study. In addition to the densely connected layers of the FNN, all LSTM RNN models begin with a LSTM class layer which initializes the use of the RNN along with the necessary short and long-term gating. Only a single LSTM layer is defined although multiple densely connected layers can be added afterward. Hyperparameters for the densely connected layers were the same as those used in the FNN models although pilot testing indicated use of more than one dense layer would lead to overfitting to the training data. Additionally, similar constraint definitions on the kernels and starting bias were found and again set as a maximized normal constraint with a maximum value of 2 for the kernel with unconstrained bias. Tuning selection instead focused on the number of nodes to retain in the LSTM layer and considered either 32 or 16 nodes.

One-Dimensional Convolutional Neural Networks

Rather than retaining varying amount of prior sequence data like the LSTM models, convolutional neural networks use a sliding window (convolution or kernel) to extract successively smaller sets of contiguous sequences which are then pooled and flattened before being processed as a one-dimensional vector through a standard densely connected network. Although primarily used for image processing and recognition for feature extraction using a two-dimensional window, one-dimensional convolutions can be used to shift unidirectionally along a single axis such as for applications using temporal data. Like the layer ordering of LSTM RNN models, 1D CNN architecture begins with a series of alternating convolution and pooling layers before a flattening layer feeds into a series of FNN dense layers. In addition to the standard layer, node, and constraint hyperparameters, CNNs require tuning for the size of the convolution window, the size of the shift the kernel takes, and the amount of sequence reduction during pooling. Given the small size of the individual temporal sequences in the current dataset, the convolution window was set at a size of two with only single steps to consider all possible adjoining sequence components. Use of a minimally defined convolution window in turn limited the networks to only a single set of convolution and pooling layers. Instead, like the LSTM, hyperparameter tuning focused on the number of nodes to pass to the final densely connected layer and again considered either 32 or 16 nodes. Constraint hyperparameters on the kernels and starting bias followed the same piloted conventions as previously described with the same parameterizations.

Prediction of Trajectories in Sequential Neural Networks

Since all neural networks were modelled using Python and TensorFlow, specific conventions were required with respect to data formats. Specifically, multi-dimensional arrays (tensors) had to be constructed as inputs to be used by TensorFlow with each row corresponding to a single observation for a participant with columns corresponding to timepoints and array slices representing individual features (outcome of interest, age, etc.). Sequential representation was done by backfilling each row with prior data with the right-most column as the value for the current observation. Since complete matrices were required for backpropagation, both CNN and RNN models began with a masking layer which defined missing values to be ignored by the networks which could then be used to apply the sequence padding in the missing cells.

Of note, predictions of neural networks in TensorFlow are iterative in nature with each row being calculated individually instead of in bulk using standard linear algebra techniques. While this is not a concern for the FNN models which only have two-dimensional arrays, or when predicting observational forecasts where prior sequences where already known and populated, it presented a problem when developing whole subject trajectories. Specifically, the outcome slices for the input arrays would consist solely of masked values and could not be used as viable inputs for the starting masking layers of the 1D CNN and LSTM models. To address this, starting values for the outcomes were seeded using the previously described baseline FNN model and used to populate the first column of the outcome slice of the response array. This starting sequence was then applied to the sequential neural networks with predictions for the next time point calculated and then used to update the response slice. This process was then repeated until all subsequent observations had been predicted.

Comparison of Model Prediction Performance

The previously described evaluation metrics for each ML model were first compared against their respective reference model before being compared pairwise against each other. This gave each model pairing four different types of comparisons which were based on a combination of both class of outcome (ADAS-Cog or CDR impairment) and type of prediction (whole subject trajectory or observation forecasting). The raw difference in evaluation metric was calculated along with percent difference relative to the poorer performing model. Statistical significance of model improvement was done by using 1000-fold bootstrapping to calculate 95% confidence intervals (CI) of the metric differences as well as one-sided proportional *p*-values to identify statistically significant model differences in prediction performance as either greater or less than zero.

Approach for Evaluation of Subject-Specific Effects

For aim 3, the role of subject-specific effects was investigated to determine how imputation or suppression of subject-specific effects impacted variance and bias when predicting both whole trajectories and observational forecasting. General modifications to the previously described methods included an emphasis on the ADAS-Cog methods to expressly investigate subject-specific effects in the CPath model. Furthermore, performance metric evaluations were limited to RMSE and AVB although the bootstrapping process for statistical assessments remained the same. In addition to the pre-specified CPath parameterization, an analogous mixed-effects beta regression (BR) model was built *de novo* from the dataset to calculate updated values for the population-level fixed effects and subject-specific random effect components as well as the MERF model described in aim 1. The meta-database was again used as the primary data source although data points were considered in 6 month increments out to only 24 months of evaluation to align with the original development of the CPath model. In addition, only the covariates used by the CPath model as previously described were used when building the *de novo*

BR (DN BR) and MERF models to allow for more equitable model comparison emphasizing the subject-specific effects without influence of feature set width.

Development of Synthetic Datasets for Validation

To complement the model evaluation for the meta-database and test generalizability of the influence of subject-specific effects, validation datasets were generated using simulation of 500 separate synthetic cohorts each with 400 participants in 6 month increments out to 60-months of evaluation. Cohorts first sampled the population-level covariates used by the CPath parameterization before generating panels of simulated ADAS-Cog subscale scores expected from subjects with equivalent demographic characteristics. As defined by the CPath model, simulated population-level covariates were baseline age, sex, *APOE4* allele counts, and baseline MMSE and were generated to create cohorts similar in disposition to the meta-database as a representative population expected for studies in cognitive decline. Ages were randomly sampled from the observed meta-database cohort with additional demographics synthetically generated using classification and regression trees to create similar marginal combinations of covariates with assistance from the `synthpop` package in R (Nowok et al., 2016). Final evaluation timepoints for each synthetic subject were randomly permuted to simulate a 15% dropout rate followed by a row-wise deletion of 15% of all remaining timepoints to simulate reasonably anticipated missingness in a real-world study.

To create the longitudinally correlated ADAS-Cog panel data, Gower's distance was first calculated among the actual subjects in the meta-database according to the population-level covariates described above. This distance was used to cluster the meta-database subjects using weighted median spheroid distance to create 20 distinct similarity clusters. Simulated participants were assigned to the nearest meta-database cluster according to their generated demographics and randomly linked to the ADAS-Cog measures of an actual meta-database subject within the same similarity cluster. A mixed-effects beta regression model for each cluster was created using these linked ADAS-Cog measures with cubic polynomial time as fixed effects with random intercepts and slopes using unstructured covariance. Each synthetic subject then had new ADAS-Cog measures generated according to their corresponding cluster-specific model with fixed and random effects randomly generated from the model covariance matrices using multivariate normal sampling. To accommodate the extended 60-month timeframe and generalization to other datasets, the covariance matrices were relaxed to allow for more varied ADAS-Cog scores at later timepoints. This process generated unique panels of ADAS-Cog scores for each simulated participant while retaining serial correlation and within-subject covariance structure expected from real-world subjects with similar population-level demographics and characteristics.

Influence of Subject-Specific Effects on ADAS-Cog

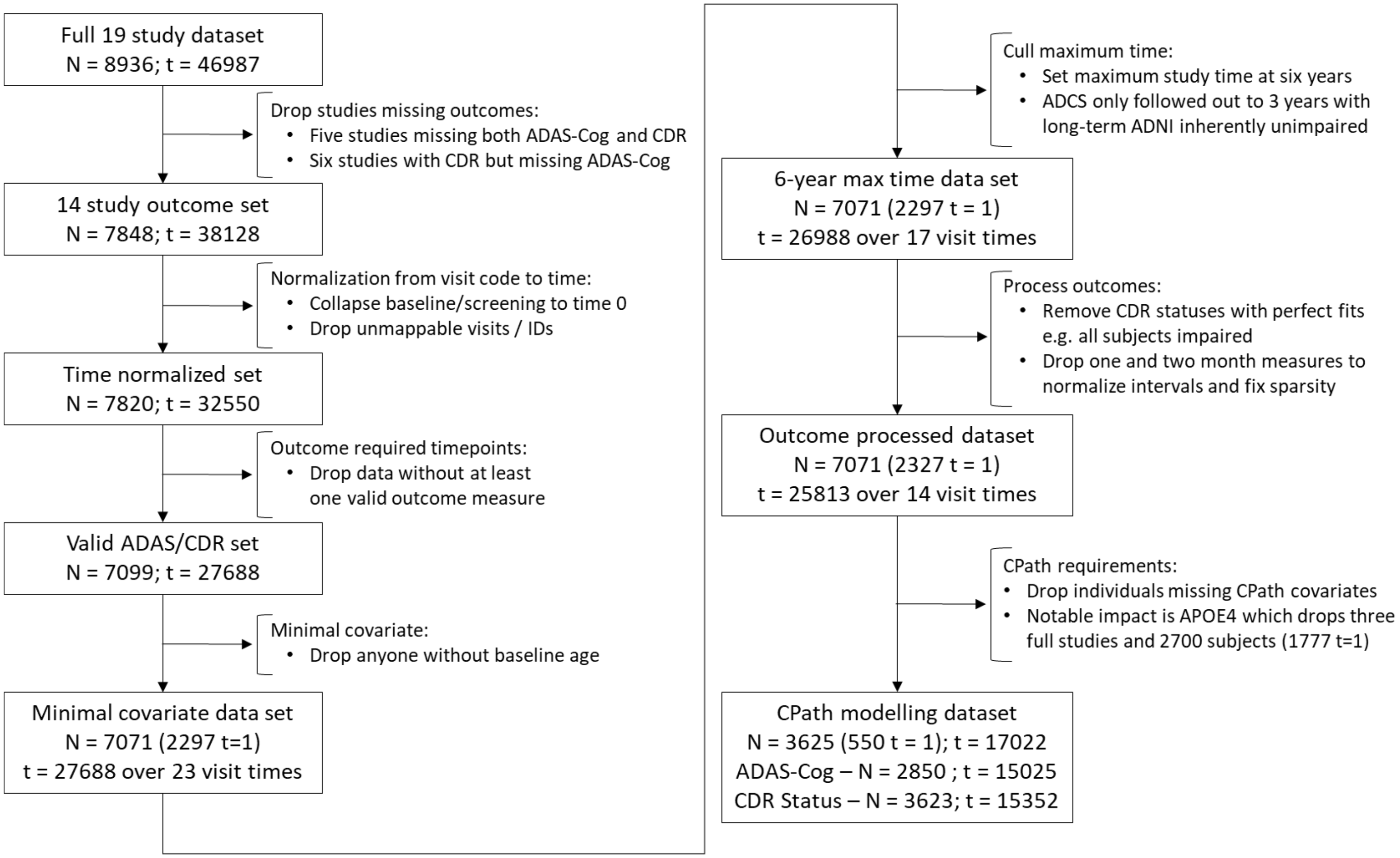
As in aims 1 and 2, evaluation was performed on holdout sets sampled from both the meta-database and the simulated data for two types of ADAS-Cog predictions: whole trajectories for subjects across all time points and forecasting of final observations. Larger validation holdouts were used for this aim but followed the same process as previously described, beginning with 20% of all subjects held out for whole trajectories followed by

20% of final observations of the model building datasets to forecast final observations. Holdout sampling was repeated 200 times for the meta-database and performed once for each of the 500 synthetic cohorts. For each model and prediction type, subject-specific effects (SSE) were either 1) suppressed with only population-level effects (PLE) used, 2) robustly imputed using 100 samplings from the random effects covariance parameters, or 3) applied directly based on the *de novo* BR and MERF model fitted values when forecasting final observations. Evaluation metric comparisons focused on pairings either within model design (CPath, DN BR, MERF) or subject-specific effect structures (population only, imputation of subject-specific effects, known fitted effects). The impact on predictive capacity for ADAS-Cog considered the differences in model performance based on the metrics of RMSE and AVB using the previously described bootstrapping procedures but also identified models with improved metrics in at least 90% of the meta-database samplings and synthetic cohorts.

TABLES AND FIGURES – METHODS AND APPROACH

Figure 1

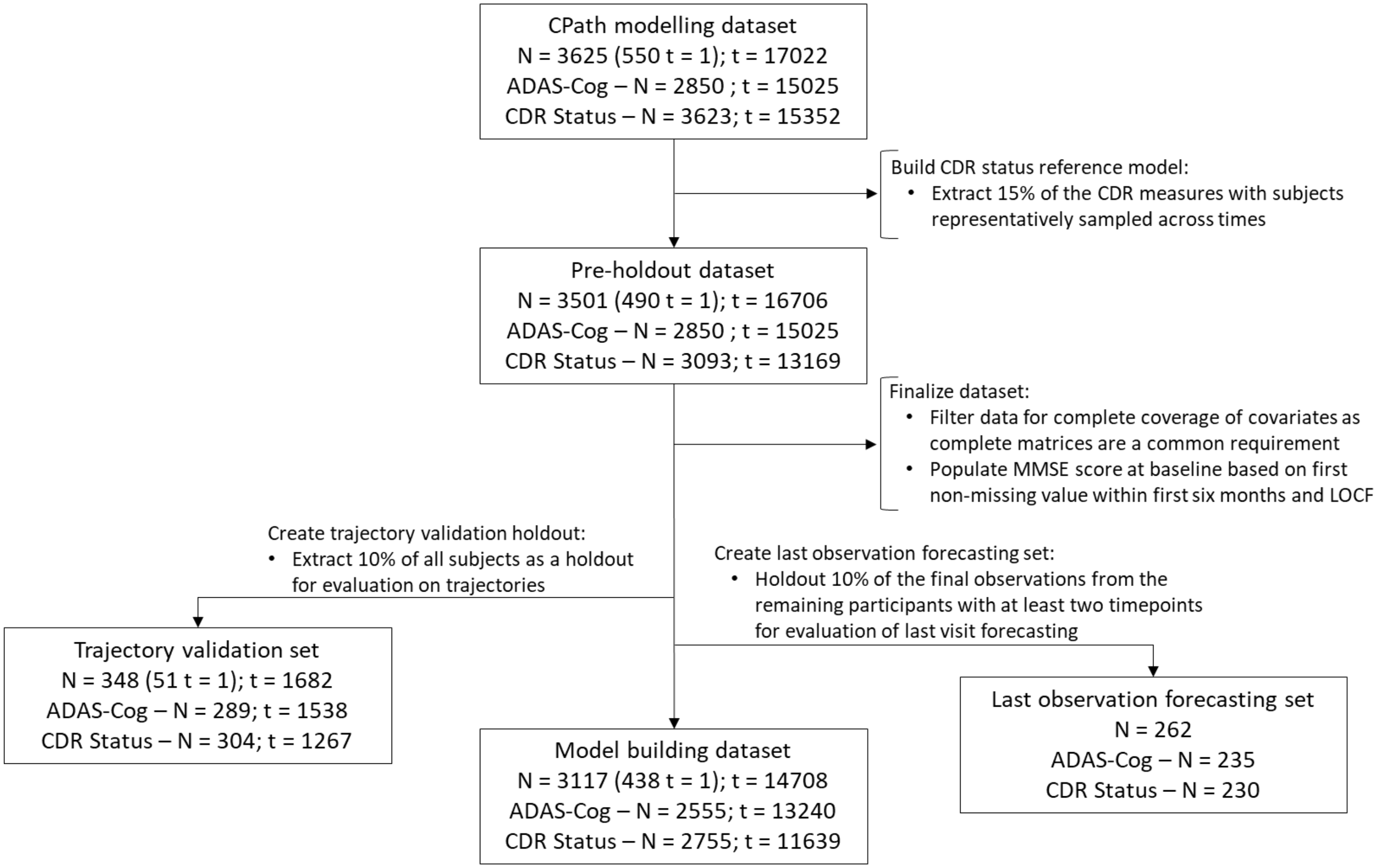
Harmonization Flow of Meta-Database Development to Pre-Holdout Dataset


Full 19 study dataset
N = 8936; t = 46987
Drop studies missing outcomes:
• Five studies missing both ADAS-Cog and CDR
• Six studies with CDR but missing ADAS-Cog
14 study outcome set
N = 7848; t = 38128
Normalization from visit code to time:
• Collapse baseline/screening to time 0
• Drop unmappable visits / IDs
Time normalized set
N = 7820; t = 32550
Outcome required timepoints:
• Drop data without at least one valid outcome measure
Valid ADAS/CDR set
N = 7099; t = 27688
Minimal covariate:
• Drop anyone without baseline age
Minimal covariate data set
N = 7071 (2297 t=1)
t = 27688 over 23 visit times
Cull maximum time:
• Set maximum study time at six years
• ADCS only followed out to 3 years with long-term ADNI inherently unimpaired
6-year max time data set
N = 7071 (2297 t = 1)
t = 26988 over 17 visit times
Process outcomes:
• Remove CDR statuses with perfect fits e.g. all subjects impaired
• Drop one and two month measures to normalize intervals and fix sparsity
Outcome processed dataset
N = 7071 (2327 t = 1)
t = 25813 over 14 visit times
CPath requirements:
• Drop individuals missing CPath covariates
• Notable impact is APOE4 which drops three full studies and 2700 subjects (1777 t=1)
CPath modelling dataset
N = 3625 (550 t = 1); t = 17022
ADAS-Cog – N = 2850 ; t = 15025
CDR Status – N = 3623; t = 15352

Note. Lists number of subjects (N), total time points (t), and subjects with only one evaluation time (t=1); including by outcome.

Figure 2

Development of CDR Reference Dataset, Training Holdout Set, and Testing Holdout Set for the Meta-Database



CPath modelling dataset
N = 3625 (550 t = 1); t = 17022
ADAS-Cog – N = 2850 ; t = 15025
CDR Status – N = 3623; t = 15352

Build CDR status reference model:

- Extract 15% of the CDR measures with subjects representatively sampled across times

Pre-holdout dataset
N = 3501 (490 t = 1); t = 16706
ADAS-Cog – N = 2850 ; t = 15025
CDR Status – N = 3093; t = 13169

Finalize dataset:

- Filter data for complete coverage of covariates as complete matrices are a common requirement
- Populate MMSE score at baseline based on first non-missing value within first six months and LOCF

Create trajectory validation holdout:

- Extract 10% of all subjects as a holdout for evaluation on trajectories

Create last observation forecasting set:

- Holdout 10% of the final observations from the remaining participants with at least two timepoints for evaluation of last visit forecasting

Trajectory validation set
N = 348 (51 t = 1); t = 1682
ADAS-Cog – N = 289; t = 1538
CDR Status – N = 304; t = 1267

Model building dataset
N = 3117 (438 t = 1); t = 14708
ADAS-Cog – N = 2555; t = 13240
CDR Status – N = 2755; t = 11639

Last observation forecasting set
N = 262
ADAS-Cog – N = 235
CDR Status – N = 230

Note. Follows from the CPath modelling dataset at the end of Figure 1 with the same conventions.

Table 1

Continuous Subject Characteristics at Baseline Within the Meta-Database

	Training dataset			Trajectory holdouts			Observation holdouts		
Baseline characteristic	All subjects N = 2755	CDR 0 subjects N = 453	CDR 0.5+ subjects N = 2302	All subjects N = 304	CDR 0 subjects N = 52	CDR 0.5+ subjects N = 252	All subjects N = 230	CDR 0 subjects N = 34	CDR 0.5+ subjects N = 196
ADAS-Cog score	12.9 ± 8.02	5.9 ± 2.90	14.5 ± 7.96	12.8 ± 8.12	6.2 ± 3.22	14.4 ± 8.15	12.5 ± 7.15	6.0 ± 2.85	13.8 ± 7.05
Age (yrs)	73.7 ± 7.58	73.8 ± 6.03	73.7 ± 7.85	74.0 ± 7.92	73.8 ± 5.48	74.1 ± 8.34	73.9 ± 7.21	73.8 ± 5.82	74.0 ± 7.43
MMSE score	25.0 ± 4.46	29.0 ± 1.22	24.3 ± 4.45	24.9 ± 4.65	29.1 ± 1.01	24.1 ± 4.64	25.6 ± 4.04	28.8 ± 1.37	25.0 ± 4.10
Weight (lbs)	144.2 ± 49.0	152.7 ± 46.8	142.5 ± 49.2	144.2 ± 49.9	154.9 ± 50.4	142.0 ± 49.7	146.3 ± 48.7	158.6 ± 45.8	144.2 ± 49.0
Systolic blood pressure	133.8 ± 17.3	134.0 ± 16.7	133.8 ± 17.4	133.0 ± 16.4	134.6 ± 16.4	132.7 ± 16.5	133.8 ± 16.7	131.7 ± 15.0	134.2 ± 17.0
Diastolic blood pressure	74.2 ± 9.78	73.6 ± 9.77	74.3 ± 9.78	74.4 ± 9.20	73.7 ± 9.07	74.5 ± 9.23	74.1 ± 10.0	72.5 ± 9.95	74.3 ± 10.0

Note. Cohorts defined according to final holdout groups as displayed in Figure 2. Data are displayed as mean ± standard deviation.

Table 2

Categorical Subject Characteristics at Baseline Within the Meta-Database

	Training dataset			Trajectory holdouts			Observation holdouts		
Baseline characteristic	All subjects N = 2755	CDR 0 subjects N = 453	CDR 0.5+ subjects N = 2302	All subjects N = 304	CDR 0 subjects N = 52	CDR 0.5+ subjects N = 252	All subjects N = 230	CDR 0 subjects N = 34	CDR 0.5+ subjects N = 196
Sex									
Male	1436 (52.1)	225 (49.7)	1211 (52.6)	160 (52.6)	22 (42.3)	138 (54.8)	119 (51.7)	20 (58.8)	99 (50.5)
Female	1319 (47.9)	228 (50.3)	1091 (47.4)	144 (47.4)	30 (57.7)	114 (45.2)	111 (48.3)	14 (41.2)	97 (49.5)
APOE4 allele counts									
Non-carriers	1266 (46.0)	314 (69.3)	952 (41.4)	151 (49.7)	42 (80.8)	109 (43.3)	116 (50.4)	22 (64.7)	94 (48.0)
Heterozygous *APOE4* carriers	1145 (41.6)	128 (28.3)	1017 (44.2)	116 (38.2)	7 (13.5)	109 (43.3)	88 (38.3)	11 (32.4)	77 (39.3)
Homozygous *APOE4* carriers	344 (12.5)	11 (2.4)	333 (14.5)	37 (12.2)	3 (5.8)	34 (13.5)	26 (11.3)	1 (2.9)	25 (12.8)
Race									
White	2519 (91.4)	413 (91.2)	2106 (91.5)	283 (93.1)	49 (94.2)	234 (92.9)	209 (90.9)	32 (94.1)	177 (90.3)
Black / African American	129 (4.7)	27 (6.0)	102 (4.4)	14 (4.6)	3 (5.8)	11 (4.4)	7 (3.0)	1 (2.9)	6 (3.1)
Other race	107 (3.9)	13 (2.9)	94 (4.1)	7 (2.3)	0 (0.0)	7 (2.8)	14 (6.1)	1 (2.9)	13 (6.6)

	Training dataset			Trajectory holdouts			Observation holdouts		
Baseline characteristic	All subjects N = 2755	CDR 0 subjects N = 453	CDR 0.5+ subjects N = 2302	All subjects N = 304	CDR 0 subjects N = 52	CDR 0.5+ subjects N = 252	All subjects N = 230	CDR 0 subjects N = 34	CDR 0.5+ subjects N = 196
Ethnicity									
Non-Hispanic / unknown	2653 (96.3)	439 (96.9)	2214 (96.2)	295 (97.0)	51 (98.1)	244 (96.8)	218 (94.8)	34 (100.0)	184 (93.9)
Hispanic	102 (3.7)	14 (3.1)	88 (3.8)	9 (3.0)	1 (1.9)	8 (3.2)	12 (5.2)	0 (0.0)	12 (6.1)
Highest education									
Less than HS	217 (7.9)	13 (2.9)	204 (8.9)	30 (9.9)	0 (0.0)	30 (11.9)	18 (7.8)	1 (2.9)	17 (8.7)
High school diploma / GED	516 (18.7)	40 (8.8)	476 (20.7)	55 (18.1)	2 (3.8)	53 (21.0)	42 (18.3)	3 (8.8)	39 (19.9)
Some college	568 (20.6)	92 (20.3)	476 (20.7)	50 (16.4)	7 (13.5)	43 (17.1)	40 (17.4)	9 (26.5)	31 (15.8)
College degree	631 (22.9)	108 (23.8)	523 (22.7)	69 (22.7)	18 (34.6)	51 (20.2)	64 (27.8)	8 (23.5)	56 (28.6)
Some post-grad	460 (16.7)	102 (22.5)	358 (15.6)	53 (17.4)	14 (26.9)	39 (15.5)	42 (18.3)	8 (23.5)	34 (17.3)
Graduate degree	363 (13.2)	98 (21.6)	265 (11.5)	47 (15.5)	11 (21.2)	36 (14.3)	24 (10.4)	5 (14.7)	19 (9.7)
Anti-dementia medication use									
No medication	1313 (47.7)	423 (93.4)	890 (38.7)	154 (50.7)	50 (96.2)	104 (41.3)	109 (47.4)	30 (88.2)	79 (40.3)
Yes medication	1442 (52.3)	30 (6.6)	1412 (61.3)	150 (49.3)	2 (3.8)	148 (58.7)	121 (52.6)	4 (11.8)	117 (59.7)

Note. Cohorts defined according to final holdout groups as displayed in Figure 2. Data are displayed as counts with percentages according to column margins in parenthesis.

RESULTS

Reference Model Results

Performance metrics for the ADAS-Cog predictions from the reference CPath model are presented for both the entire testing dataset as well as at individual time points in Table 3 for whole subject trajectories with scatterplots of the true and predicted values as well as plots of overestimation and underestimation found in Figures 3 and Figure 4 respectively. Equivalent performance metrics for final observation forecasts are shown in Table 4 with prediction and discrepancy scatterplots in Figures 5 and 6. For the entire dataset, predictions errors for both RMSE and MAE were numerically higher for forecasting compared to trajectories. As expected, RMSE values were larger than MAE values due to especially large ADAS-Cog scores but the relative increases were not notably different between trajectories (RMSE: 6.82; MAE: 5.28, 30% increase) and forecasts (RMSE: 9.72; MAE: 6.86, 40% increase). Biases indicated a general overestimation of ADAS-Cog scores for trajectories, largely due to an inability of the CPath model to accurately predict observed scores in the lower range of 0 to 5. Conversely, forecast observations were generally underestimated with an inability to accurately predict especially large ADAS-Cog values. Of interest, AVB values were rather similar between trajectories and forecasts and biases were numerically similar with differences predominately in the sign. Importantly, these error and bias results indicate relatively similar spread in ADAS-Cog scores between the two holdout testing sets.

Performance metrics for the CDR impairment predictions from the reference logistic mixed model classifier are presented for both the entire testing dataset as well as at individual time points in Table 5 for whole subject trajectories with bar charts comparing the counts of observed and predicted impaired meta-database subjects as well as the proportion of false positives and false negatives marginalized on time found in Figures 7 and 8. Equivalent performance metrics for final observation forecasts are shown in Table 6 with prediction and misclassification bar plots in Figures 9 and 10. Compared to the predictions for ADAS-Cog from the CPath model, performances of the reference logistic classifier were more similar between trajectories of CDR impairment and final observation forecasts. Accuracy was high at over 70% with especially high precisions of over 95%, a somewhat expected result due to the imbalance of CDR status in the training data set with 80.8% of participants with CDR scores of 0.5 or greater. Both types of predictions had a tendency towards false negatives as indicated by comparatively lower recalls of 66% for trajectories and 69% for forecasts. ROC AUC values were relatively large for both types of predictions indicating good overall classification ability by the reference classifiers even without considerations of optimal cut points.

Prediction from Ensemble Methods

ADAS-Cog Results

Summary sets of the evaluation metrics for the ADAS-Cog predictions across all timepoints for both whole subject trajectories as well as final observation forecasts for the ensemble methods of aim 1 can be found in Table 7. Results of the cross-fold validation sweeps conducted during hyperparameter selection for the MERF, bagged GLMM trees, and boosted mixed-effects trees can be found in Figures 11-13 with results of the top performing hyperparameter combinations for each method in Table 8.

When optimizing the MERF models on ADAS-Cog, hyperparameter evaluation was unable to reach consensus, with each metric implicating a different configuration set. However, there was relatively little difference between the hyperparameter combinations although utilization of only 20% of the feature space was uniformly associated with larger prediction error. The final hyperparameter set selected used 40% of the feature space during sampling and constructed 750 trees per forest.

For the MERF models, performance metrics for whole subject trajectories are in Table 9 with overall statistical evaluation of the prediction metrics compared to the CPath reference model in Table 10. For whole subject trajectories, scatterplots of predicted and actual ADAS-Cog scores and the corresponding estimation discrepancies are visualized in Figures 14 and 15. For final observation forecasts, equivalent performance metric and evaluation results can be found in Tables 11 and 12 with scatterplots of predicted and actual scores along with estimation discrepancies shown in Figures 16 and 17. All metrics indicated improved performance of the MERF model relative to the CPath reference for both whole subject trajectories and observation forecasts. Unlike the reference model, prediction errors were similar between the two types of predictions with slightly smaller RMSE (4.93 versus 4.72) and MAE (3.79 versus 3.59) values for forecasts. Raw prediction bias was notably small across the full dataset for trajectories with a value of 0.07 although this was largely due to averaging as bias values tended to underestimate at earlier timepoints and overestimate at later times with a similar pattern for forecasts. This was also reflected with the AVB values which were again smaller for forecasts (2.82) compared to trajectories (3.16).

Optimization for the bagged GLMM trees was able to reach consensus among the four considered configurations, with a combination of only 100 trees using a larger sample fraction size of 75% being the preferred design. Differences in performance were much more dependent on the sample size fraction instead of the number of trees used during bagging, with relatively little difference between 100 and 200 trees within each of the two fraction sets but improvements were observed in all metrics when larger sample fractions were used.

For the single GLMM tree models, performance metrics for whole subject trajectories are in Table 13 with overall statistical evaluation of the prediction metrics compared to the CPath reference model in Table 14. For whole subject trajectories, scatterplots of predicted and actual ADAS-Cog scores and the corresponding estimation discrepancies are visualized in Figures 18 and 19. For final observation forecasts, equivalent performance metric and evaluation results can be found in Tables 15 and 16 with scatterplots of predicted and actual scores along with estimation discrepancies shown in Figures 20 and 21. Similar results for the bagged GLMM models are presented in Tables 17 and 18 with plots in Figures 22 and 23 for whole subject trajectories as well as Tables 19 and 20 and Figures 24 and 25 for final observation forecasts. Much like the MERF model, the single GLMM tree and the bagged forests both gave superior performance compared to the CPath reference for both types of prediction errors as well as raw bias and AVB. Of note, when predicting whole subject trajectories, values for performance metrics were numerically lower for the bagged GLMMs compared to the single tree; however, the metrics were similar enough that bootstrapping during cross-model prediction did not indicate any significant difference between the two GLMM methods. Conversely, all prediction errors and biases were numerically lower for the single GLMM tree when forecasting future observations, and in the case of RMSE, mean AE and AV bias, the 95% bootstrapped confidence intervals of the differences in performance metrics between the two GLMM tree methods did not indicate statistical significance (RMSE: [-1.17, -0.03]; MAE: [-0.86, -0.10]; AVB: [-0.55, -0.01]).

As mentioned in the methods, hyperparameter selection for the boosted mixed-effects trees method for the final meta-database focused solely on the number of trees with the other parameters piloted using smaller datasets for tractability purposes with hyperparameter selection using 5-fold cross-validation instead of 10-fold as with the other models. The processing time and memory footprints for the boosted trees made computation especially challenging and use of the restricted hyperparameter configurations with a maximum tree depth of 15, a minimum of 20 observations in the terminal nodes, with a boosting shrinkage rate of 0.01 were all found to be adequate for the meta-database and its current feature space. Although there was relatively little difference in bias based on number of trees in the iterative boosting, there were appreciable reductions in both RMSE and MAE with larger trees. Consensus was reached for the boosted method with all metrics recommending the use of 200 trees over 100.

For the final boosted tree models, performance metrics for whole subject trajectories are in Table 21 with overall statistical evaluation of the prediction metrics compared to the CPath reference model in Table 22. For whole subject trajectories, scatterplots of predicted and actual ADAS-Cog scores and the corresponding estimation discrepancies are visualized in Figures 26 and 27. For final observation forecasts, equivalent performance metric and evaluation results can be found in Tables 23 and 24 with scatterplots of predicted and actual scores along with estimation discrepancies shown in Figures 28 and 29. As with the other ensemble methods, performance was superior for the boosted mixed-effects trees with reductions in RMSE, MAE and both bias calculations for both whole subject trajectories and observation forecasts. Unlike other ensemble methods, boosting had lower prediction errors when determining trajectories compared to forecasts for both RMSE (trajectories: 4.94; forecasts: 5.23) and MAE (trajectories: 3.86; forecasts: 3.97) although these were still much more numerically similar compared to the discrepancies seen in the CPath reference model. Much like the MERF model, raw bias values were smaller for trajectories compared to forecasts but again this result appeared largely driven by underestimation at earlier timepoints and overestimations further out. Also, like the MERF model, AVB values were similar between the two types of predictions although they were slightly lower for forecasts.

CDR-Based Impairment Results

Evaluation metrics for the CDR-based impairment status predictions across all timepoints for both whole subject trajectories and final observation forecasts for the ensemble methods of aim 1 can be found in Table 25. Results of the cross-fold validation sweeps conducted during hyperparameter selection for the MERF, bagged GLMM trees, and boosted mixed-effects trees can be found in Figures 30-32 with results of the top performing hyperparameter combinations in Table 26.

The most notable outcome from the CDR impairment prediction portion of the study was the exceptionally large performance metrics when forecasting, with the sequential models of aim 1 and aim 2 both showing near perfect prediction capacity with many metrics valued at 0.95 and higher. This largely seems to be a consequence of relative stability of the impairment classifier in the meta-database as CDR exhibited little change within subjects, especially for the clinical trials of the ADCS component. In particular, this highlights these model's utility and reliance on subject-specific effects. These results do not necessarily detract from the research questions at hand or model comparison pipeline but should temper the interpretation of any forecasting results.

When optimizing the MERF models for CDR classification, consensus was not unanimous across the four considered configuration sets but did indicate general better model performance with more trees and larger proportions of feature subsets. Similar to the ADAS-Cog tuning, improvements from larger feature sampling proportions were more sizable compared to performance improvements from simply increasing the number of trees in the random forests. Recall, in particular, observed additional benefit from feature space increases. Due to the general similarity across configurations for precision and ROC AUC combined with the previously noted improvements in recall, the final hyperparameter configuration for the MERF classifier was 750 trees per random forest with sampling of 70% of the feature space.

For the MERF models, performance metrics for whole subject trajectories are in Table 27 with overall statistical evaluation of the prediction metrics compared to the logistic classifier reference model in Table 28. For whole subject trajectories, bar plots of the predicted and actual impaired subject counts along with the proportion of misclassifications within each timepoint can be found in Figures 33 and 34. For final observation forecasts, equivalent performance metric and evaluation results can be found in Tables 29 and 30 with bar plots of predicted and actual impaired counts along with misclassification rates by timepoint shown in Figures 35 and 36. As when predicting ADAS-Cog scores, the MERF model exhibited significantly improved performance when classifying CDR impairment compared to the logistic reference model for both whole trajectories and observation forecasting. Precision was numerically higher although bootstrapping did not indicate statistically significant increases from the already high precision of the reference design. However, accuracy, despite imbalance, and AUC were larger for both types of predictions and recall saw notable gains with a 20.8% improvement in recall for whole subject trajectories and a 40.8% improvement for forecasts. This indicates a reduction in false negative classification for the MERF models compared to the reference and was reflected in the associated NRI values of 0.124 and 0.285 for trajectories and forecasts respectively, which highlight the improved reclassification of the MERF model.

Hyperparameter tuning for the bagged GLMM trees on CDR impairment status did not reach unanimous consensus but did give general suggestions. Recall and accuracy both saw the greatest gains when larger proportions of the subject dataspace were sampled during bagging while precision and ROC AUC, although reduced with the larger proportions, were much more numerically similar in comparison. At these higher bagging proportions, the number of trees was a much less important consideration although nearly all metrics implicated the use of fewer trees. Based on these results, the final configuration design used a 0.75 proportion of the subjects during sampling with 100 trees for aggregation.

For the single GLMM models, performance metrics for whole subject trajectories are in Table 31 with overall statistical evaluation of the prediction metrics compared to the logistic classifier reference model in Table 32. For whole subject trajectories, bar plots of the predicted and actual impaired subject counts along with the proportion of misclassifications within each timepoint can be found in Figures 37 and 38. For final observation forecasts, equivalent performance metric and evaluation results can be found in Tables 33 and 34 with bar plots of predicted and actual impaired counts along with misclassification rates by timepoint shown in Figures 39 and 40. Similar tables and plots for the bagged GLMM model are in Tables 35 and 36 and Figures 41 and 42 for whole subject trajectories as well as Tables 37 and 38 and Figures 43 and 44 for final observation forecasts. The same patterns of classification improvement observed with the MERF model were also seen with the single GLMM tree model and bagged GLMM trees. Numerically, all evaluation metrics were higher for both trajectories and forecasts in both types of GLMM designs with all being significantly greater than the logistic classifier, with the exception of precision, which as mentioned was already high in the reference model. Reductions in false negatives were sizable with improvements in trajectory recall of 17.5% and 21.3% and forecasting recall of 39.9% and 39.2% for the single GLMM and bagged trees respectively. Also, like the MERF model, the improvements in classification, while sizable for whole subject trajectories, where exceptionally great when applied to observation forecasting, highlighting the role of prior knowledge when evaluating the especially stable impairment classification. For example, reclassification improvement in trajectories as indicated by NRI was 0.128 and 0.127 for the single tree and the bagged trees respectively but 0.285 and 0.302 for final observation forecasting. A noteworthy difference for the GLMM classifiers was that the improvements in forecasting previously observed in the single GLMM tree under a regression framework for prediction of ADAS-Cog were no longer observed for CDR classification. Numerically, many evaluation metrics were better under the bagged framework for both trajectories and forecast predictions such as the ROC AUC (single GLMM trajectory: 0.892; bagged trajectory: 0.901; single GLMM forecast: 0.981; bagged forecast: 0.987); however, the only difference between GLMM models considered statistically significant was recall for trajectories (95% CI: [0.001, 0.050], *p*=0.021) with a mild trend for the ROC AUC of whole-subject trajectories (95% CI [-0.003, 0.031], *p*=0.056).

Optimization of the boosted mixed-effects trees again only considered the number of trees in the boosting sequence with the same configuration for tree depth, terminal node observations, and shrinkage rate using 5-fold cross-validation for configuration tuning. When contrasted to the tuning for ADAS-Cog, differences in metrics between 100 trees and 200 trees were much smaller and unanimous consensus was not reached. Recall and accuracy gave slight preference to 100 trees while AUC and precision gave priority to 200 boosting trees. Since the improvements in recall for the boosted trees were generally smaller compared to the equivalent tuning improvements observed in the other ensemble classifiers, the larger set of 200 boosted trees was once again selected for impairment classification using CDR.

For the boosted trees, performance metrics for whole subject trajectories are in Table 39 with overall statistical evaluation of the prediction metrics compared to the logistic classifier reference model in Table 40. For whole subject trajectories, bar plots of the predicted and actual impaired subject counts along with the proportion of misclassifications within each timepoint can be found in Figures 45 and 46. For final observation forecasts, equivalent performance metric and evaluation results can be found in Tables 41 and 42 with bar plots of predicted and actual impaired counts along with misclassification rates by timepoint shown in Figures 47 and 48. The same pattern observed in all previous ensemble classifiers was once again seen for the boosted trees design. Improvements were seen by all metrics for both types of predictions when compared to the logistic mixed-effects reference classifier and were considered statistically significant by bootstrapping, except for precision. Once again, the most sizable increases were in false negative reduction with recall increasing by 22.0% for trajectories and 35.2% for forecasts. This was reflected in the NRI increases compared to the reference classifier with reclassification values of 0.147 and 0.246 respectively under the new designs. However, the computational performance of the boosted trees in terms of build time and memory demands was once more the poorest compared to all other ML models and, in fact, was even worse for CDR impairment classification when compared to the computational demands when modelling the ADAS-Cog with boosted trees. This is even after the previously described adjustments in model building by running the models in parallel with fewer cross-folds for model tuning. Although the boosted trees design still performed well, for example, it displayed the highest ROC AUC for all models for trajectories with a value of 0.916, its predictive performance was arguably not powerful enough to justify the additional requirements in time and computational resources as well as implementation modifications necessary for this particular ensemble method.

Prediction from Neural Networks

ADAS-Cog Results

Evaluation metrics for the ADAS-Cog predictions across all timepoints for both whole subject trajectories and final observation forecasts for the various neural networks of aim 2 can be found in Table 43. Results of the cross-fold validation sweeps conducted during hyperparameter selection for the feed-forward neural networks, one-dimensional convolutional neural networks, and long short-term memory recurrent neural networks can be found in Figures 49-51 with results of the top performing hyperparameter combinations in Table 44.

Hyperparameter tuning of the FNN models generally suggested the use of three hidden layers beginning with 8 nodes although RMSE implicated the use of only two layers. The improvement in raw bias with the additional layer was substantial enough compared to the minimal differences in prediction error between configurations that the deeper model was selected even though hyperparameter consensus was not unanimous.

For the feed-forward neural network reference designs, performance metrics for whole subject trajectories are in Table 45 with overall statistical evaluation of the prediction metrics compared to the CPath reference model in Table 46. For whole subject trajectories, scatterplots of predicted and actual ADAS-Cog scores and the corresponding estimation discrepancies are visualized in Figures 52 and 53. For final observation forecasts, equivalent performance metric and evaluation results can be found in Tables 47 and

48 with scatterplots of predicted and actual scores along with estimation discrepancies shown in Figures 54 and 55. Much like the ensemble methods of aim 1, the neural network designs of aim 2 showed significant improvement when predicting ADAS-Cog scores for both whole subject trajectories and final observation forecasts. This includes the FNN reference network even though it did not leverage sequential data in the same fashion as the other neural network designs. Prediction errors were much less for the FNN design compared to the CPath reference with improvements of 31.7% for RMSE and 32.6% for MAE. Notably, the FNN model demonstrated increased prediction errors for forecasting when compared to trajectory predictions, a similar pattern also observed in the CPath reference model and ostensibly due to neither model making direct use of prior sequence data when forecasting. This pattern also held with bias in the FNN models with reductions in both raw bias and AVB compared to the CPath reference which were significant under bootstrapping, but with increased bias for forecasts relative to trajectory determinations.

As described in the hyperparameter methods for the 1D CNN model, use of a minimally sized convolution kernel limited the number of convolution-pooling layers to a single set, with tuning instead focusing on the number of nodes in the convolution layer to be passed to final dense connection scoring layer considering either 32 or 16 nodes. Although bias showed greater improvement with a larger number of nodes, this was the only metric to suggest so with all prediction errors larger under the 32 node design. This was suggestive of overfitting to the training data, so consensus assigned the final hyperparameter set to the 16 node configuration.

For the 1D convolutional neural network, performance metrics for whole subject trajectories are in Table 49 with overall statistical evaluation of the prediction metrics compared to the CPath reference model in Table 50. For whole subject trajectories, scatterplots of predicted and actual ADAS-Cog scores and the corresponding estimation discrepancies are visualized in Figures 56 and 57. For final observation forecasts, equivalent performance metric and evaluation results can be found in Tables 51 and 52 with scatterplots of predicted and actual scores along with estimation discrepancies shown in Figures 58 and 59. The 1D CNN, like other ML models, was better at ADAS-Cog prediction for both trajectories and forecasting when compared to the CPath reference although not necessarily to the same degree as other ML designs. Prediction errors for MAE and absolute value of the bias were reduced for both types of predictions despite a hyperparameter configuration that was less optimized for bias. One difference of note is the 1D CNN followed the same raw bias pattern as the CPath reference with overestimation for trajectories, especially at later timepoints, and underestimation for forecasts. Meanwhile most models had similar signs for their raw bias metrics. However, the most notable effect of ADAS-Cog prediction for the 1D CNN was RMSE for trajectories. Although RMSE was numerically lower for the 1D CNN design, it was only a 4.6% reduction (1D CNN: 6.506; CPath: 6.82) and was the only ML model in both aims not to be observed as statistically different from the reference model with bootstrapping (95% CI: [-0.705, 0.050]). However, when observed data were used in the 1D CNN design, RMSE observed significant improvements with a reduction of 39.8% compared to the CPath method (95% CI: [-4.82, -2.84]). In addition, unlike the FNN reference network which did not make direct use of the sequential nature of the panel data, evaluation metrics for observation forecasting were superior compared to the metrics of trajectories, the pattern commonly observed in the sequential ensemble methods.

Optimization of the long short-term memory recurrent neural network followed the same process as the 1D CNN with tuning limited to the number of nodes in the LSTM layer. A similar pattern arose with reduced bias with a greater number of nodes at the expense of increased RMSE and MAE. Again, due to concerns with potential overfitting to the training data and loss of generalization, the smaller node set of 16 was selected for evaluation on the holdout testing set for ADAS-Cog.

For the LSTM recurrent neural network, performance metrics for whole subject trajectories are in Table 53 with overall statistical evaluation of the prediction metrics compared to the CPath reference model in Table 54. For whole subject trajectories, scatterplots of predicted and actual ADAS-Cog scores and the corresponding estimation discrepancies are visualized in Figures 60 and 61. For final observation forecasts, equivalent performance metric and evaluation results can be found in Tables 55 and 56 with scatterplots of predicted and actual scores along with estimation discrepancies shown in Figures 62 and 63. The results of performance improvement for the LSTM model were generally similar to the results of the CNN model when predicting ADAS-Cog scores. Both prediction errors and biases were reduced for the LSTM RNN including a statistically significant reduction in RMSE for whole subject trajectories (9.6% reduction, 95% CI [-0.996, -0.260]) which was not seen in the 1D CNN framework. Raw bias was numerically lower, although not significant compared to the CPath reference for either prediction type and followed the same pattern seen in the other sequential neural network design with overestimation of trajectories and underestimation of observation forecasts.

Most strikingly, the benefit of leveraging previously known observations was again apparent for the sequential RNN with improved prediction performance for both errors and biases when using previously known data for forecasting as when compared to generation of whole subject trajectories. Like the 1D CNN, this indicates the utility of the sequential neural networks when compared to the standard non-sequential FNN under a forecasting paradigm for these continuous cognitive outcomes.

CDR-Based Impairment Results

Evaluation metrics for the CDR-based impairment status predictions across all timepoints for both whole subject trajectories and final observation forecasts for the neural network methods of aim 2 can be found in Table 57. Results of the cross-fold validation sweeps conducted during hyperparameter selection for the FNN, 1D CNN, and LSTM RNN models can be found in Figures 64-66 with results of the top performing hyperparameter combinations in Table 58.

As in aim 1, the most notable result of the impairment prediction for the neural networks was the exceedingly large metrics for the two sequential networks while forecasting. That these increases were not seen to the same degree in the non-sequential FNN model gives further credence to the notion these performance metrics are being driven by the relative stability of CDR score and, by association, the corresponding impairment classification within individual subjects and how the longitudinal models are leveraging this prior knowledge.

Hyperparameter optimization for the FNN models followed the same general pattern as the ensemble methods, with relatively little difference between the two layer and three layer configurations. Unanimous consensus was not reached on these models with recall and accuracy favoring the deeper designs while precision and ROC AUC favored the shallower two layer framework. As with other models that met with this pattern, recall improvements were slightly larger compared to improvements in AUC and the desire to reduce false negatives common in the classification models led to selection of the deeper three layer design as the evaluation network.

For the FNN models, performance metrics for whole subject trajectories are in Table 59 with overall statistical evaluation of the prediction metrics compared to the logistic classifier reference model in Table 60. For whole subject trajectories, bar plots of the predicted and actual impaired subject counts along with the proportion of misclassifications within each timepoint can be found in Figures 67 and 68. For final observation forecasts, equivalent performance metric and evaluation results can be found in Tables 61 and 62 with bar plots of predicted and actual impaired counts along with misclassification rates by timepoint shown in Figures 69 and 70. As with other ML and ensemble designs, the FNN model showed improvement over the mixed model logistic classifier across all performance metrics. All classification improvements were also statistically significant with the exception of precision, which was known to be high even for the standard inferential model. Unlike when the FNN model predicted ADAS-Cog scores, the classification metrics for observation forecasting were improved when compared to the performance for the whole subject trajectories. However, these improvements were much more mild when compared to results observed in the sequential ANN methods or the ensemble classifiers of aim 1, most likely for the CDR stability reasons and use of known sequential data aspects cited above. This was reflected with similar NRI values for trajectories and forecasts of 0.134 and 0.123 respectively, with the decrease for observation forecasts largely due to a preponderance of false negatives at later times.

Hyperparameter optimization for the 1D CNN and LSTM RNN led to very similar results despite the difference in architecture of the networks. Consensus for the 1D CNN was unanimous with all metrics indicating improved performance for the smaller convolutional layer using only 16 nodes instead of 32. Nearly all metrics for the RNN also recommended the smaller LSTM layer of 16 nodes with the exception of recall, which suggested a wider layer of 32 nodes. However, the increase in recall was smaller than the increases in the other classification metrics so the final configuration utilized 16 nodes in its LSTM layer. These tuning results were the same as those for the ADAS-Cog with increased node counts generally leading to overfitting to the training data given the relatively small size of the feature space for the meta-database.

For the 1D CNN, performance metrics for whole subject trajectories are in Table 63 with overall statistical evaluation of the prediction metrics compared to the logistic classifier reference model in Table 64. For whole subject trajectories, bar plots of the predicted and actual impaired subject counts along with the proportion of misclassifications within each timepoint can be found in Figures 71 and 72. For final observation forecasts, equivalent performance metric and evaluation results can be found in Tables 65 and 66 with bar plots of predicted and actual impaired counts along with misclassification rates by timepoint shown in Figures 73 and 74. Improvements in the CDR impairment classification of the 1D CNN were much better compared to the reference classifier for both types of predictions. All metrics were greater for the sequential NN and were statistically significant by bootstrapping except for precision on whole subject trajectories. The added utility of prior sequence by these sequential models was again displayed with vast improvements in metrics during observation forecasting compared to prediction of whole subject trajectories (trajectory NRI: 0.159; forecast NRI: 0.787). As with other sequential designs, false negative reduction was the most apparent improvement for both prediction types. Although there was a somewhat higher than expected false negative rate when forecasting impairment at later time points, this was still minimal compared to the reference logistic classifier with a recall of 0.944 across all data points. However, forecasting also saw a relatively high precision of 0.994 indicating a high capacity to identify impairment when compared to the other sequential models. A final observation was that although the trajectory performance for the 1D CNN was not markedly better than the FNN control for all metrics, the relative improvements compared to the reference model were not as stark as observed with the ADAS-Cog, with recall and accuracy both higher for the 1D CNN compared to the FNN.

For the LSTM RNN, performance metrics for whole subject trajectories are in Table 67 with overall statistical evaluation of the prediction metrics compared to the logistic classifier reference model in Table 68. For whole subject trajectories, bar plots of the predicted and actual impaired subject counts along with the proportion of misclassifications within each timepoint can be found in Figures 75 and 76. For final observation forecasts, equivalent performance metric and evaluation results can be found in Tables 69 and 70 with bar plots of predicted and actual impaired counts along with misclassification rates by timepoint shown in Figures 77 and 78. As with all other ML models, the LSTM neural network was a much better classifier of CDR impairment when compared to the logistic mixed-effects model. Precision was numerically higher for both trajectories and forecasts, although not significant under bootstrapping, with all other performance metrics statistically greater than the logistic reference model framework. False negatives were especially reduced with recalls of 0.824 for trajectories and 0.983 for forecasts, both of which were the highest for all models in both aims. As with the 1D CNN, the reduction in metrics for trajectories relative to the FNN reference network were not as pronounced for CDR impairment classification when compared to ADAS-Cog score predictions performance, with the LSTM model again showing better accuracy and recall compared to the non-sequential NN. The exceptional capacity for forecasting in this stable dataset was again on display for the sequential neural network with NRI values of 0.146 for trajectories but 0.769 for forecasting of final observations.

Cross-Model Comparisons

Comprehensive evaluations comparing the performance metrics for all machine learning models from aims 1 and 2 along with their corresponding reference model are provided for both whole subject trajectories and final observation forecasting for each combination of performance metric and type of prediction. Values are presented as numeric difference in metric between the models, percent difference in metrics, the one-sided bootstrap proportional *p*-value, and the bootstrapped 95% confidence intervals for the numeric differences. For the ADAS-Cog score predictions, comparisons for RMSE, MAE, raw bias, and AVB can be found in Tables 71-78, each for whole subject trajectories and final observation forecasts. For the classification of CDR impairment status, model comparisons for accuracy, precision, recall, and ROC AUC can be found in Tables 79-86, again within both whole subject trajectories and final observation forecasts.

As previously described, the ensemble methods of aim 1 and neural networks of aim 2 all showed significant improvements in prediction performance when compared to the inferential reference methods for both ADAS-Cog score and impairment status based on CDR, whether predicting whole subject trajectories or forecasting final observations. However, there was no model which uniformly demonstrated improved performance across all evaluated frameworks, with certain models exhibiting better performance depending on prediction type, evaluation metric, and class of outcome.

When predicting whole subject trajectories for the ADAS-Cog, the MERF and FNN models gave the best predictions with the lowest RMSE and MAE prediction errors as well as the lowest AV biases compared to the other ensemble methods and neural networks. The FNN model outperformed all other neural networks under bootstrapping while the MERF model was numerically better than the other ensemble methods, although the prediction errors and biases compared to the boosted mixed-effects trees were not significantly different. Furthermore, the evaluation metrics of the FNN model also indicated superior performance when compared to the MERF model for RMSE (95% CI: [-0.469, -0.062]), MAE (95% CI: [-0.374, -0.078]), and AVB (95% CI: [-0.478, -0.161]).

Although the FNN model predicted whole trajectories for ADAS-Cog scores well, it was not as effective at forecasting compared to the sequential ML methods. The ensemble methods had uniformly lower RMSE values compared to the neural networks with the single GLMM tree and MERF models demonstrating the lowest RMSE and MAE prediction errors compared to the other methods for ADAS-Cog forecasting. RMSE prediction errors for the neural networks were the highest observed among the models with the FNN model having numerically lower, but not statistically different, RMSE compared to the sequential CNN and RNN designs. However, MAE prediction errors were numerically better for the LSTM and CNN models compared to the bagged GLMM and boosted trees, although not significantly superior. Combined with the RMSE results, this suggests the sequential neural network models may be less capable of predicting especially extreme ADAS-Cog scores indicating a sensitivity to outlier measures. In contrast to the prediction error results, the reductions in AVB for the two sequential NN methods were pronounced, with significantly reduced bias compared to all other models except for the single GLMM tree. As a result, model selection when forecasting final ADAS-Cog observations is especially dependent on which metric is to be optimized.

The different ML models also had varied patterns of performance metrics when predicting whole trajectories of CDR impairment. Precision was especially high, with relatively few false positives for all designs using the data-driven cut point selection described in the methods. Although all models had at least as much positive predictive value as the logistic mixed-effects classifier, only the non-sequential FNN model had statistically higher precision (95% CI: [0.002, 0.025]) compared to the reference model. Additionally, the precision of the FNN model was also statistically improved compared to most other ML models. However, it must be noted precision in whole subject trajectories was high for all models with the lowest value seen in the LSTM RNN with 0.958 as its observed value. Although the FNN model had high performance for the ADAS-Cog trajectories, it did not perform as well when predicting trajectories for CDR impairment. It had the lowest accuracy and recall of all models, but this may be due to cut point optimization; a facet that may explain the increases in precision and a relatively high ROC AUC, although it was not statistically distinct from the ensemble methods. Accuracy and recall were highest for the two sequential ANN models although improvements over the boosted trees and bagged GLMM models were not statistically significant. When considering the ROC AUC metric which is agnostic to a classification cut point, the CNN and LSTM NN models were the poorest performers, albeit superior to the reference classifier, with the highest AUC observed by the boosted mixed-effects trees with a value of 0.916 which was statistically better than nearly every other model.

When evaluating the model performance for forecasting final CDR impairment status, a variety of patterns were seen, much like trajectory evaluation. As has been mentioned, the most notable feature was the exceptional improvement in predictive capacity for the sequential methods when contrasted to generation of whole subject trajectories. While this is most certainly a consequence of the disposition of the meta-database, it does still enable cross-model comparison between the sequential methods. The results most clearly supporting the role of prior sequence leveraging can be seen by noting how the non-sequential FNN model had the poorest forecasting performance across all metrics. However, the FNN model's improvement over the reference mixed model classifier indicates stable impairment class and prior data sequence does not explain all of the improved performance and implicating other facets specific to the network itself.

Among the sequential models, accuracy especially favored the LSTM RNN with a value of 0.972 using the cut point optimization; however, all sequential models had forecast accuracies of greater than 0.930, all of which outstripped both the reference (0.730) and FNN models (0.832). Precision was exceptionally high for all models, including the reference design and non-sequential FNN. In fact, the only improvement observed to be statistically significant was from the 1D CNN model compared to the logistic classifier

(95% CI: [0.005, 0.023]). Like accuracy, recall favored the LSTM RNN with a value of 0.983 but all sequential ML models had recalls of 0.940 or greater indicating exceptional reduction in false negative misclassification while using prior sequence information when compared to the reference and FNN designs. These exceptionally high values for precision, recall and accuracy were also carried over into evaluation of the ROC AUC measures. Notably, the cut point agnostic AUC metric favored the ensemble methods over the neural network designs, again showing the dependence of the other metrics on cut point selection. The highest AUC was seen in the bagged GLMM design with a value of 0.987 although all sequential models had AUCs of 0.950 or greater. Again, the sequential designs showed their specific utility for forecasting on this particular classifier when compared to the non-sequential FNN, even when acknowledging the stability of impairment classification. Although the ROC AUC for the FNN model was an improvement over the reference control with a value of 0.924 (95% CI: [0.058, 0.131]), an especially important result which highlights machine learning utility specifically, it was significantly lower compared to the other models which were able to adequately leverage previously observed data and acknowledge the limited within-subject change of CDR-based impairment.

Influence of Subject-Specific Effects on ADAS-Cog

Meta-Database Results

Summary evaluation results of RMSE, MAE and AVB for the meta-database across the 200 samplings are presented for both prediction methods with average performance metrics reported in Table 87. For whole subject trajectories, percent differences in

RMSE and AVB between models and the corresponding 95% confidence intervals presented in Table 88 with RMSE values visualized in Figure 79 and AVB values in Figure 80. Similar cross-design comparisons for final observation forecasts, including models with fitted subject-specific effects, are presented in Table 89 and visualized for RMSE and AVB in Figures 81 and 82 respectively. Within models, the imputation of subject-specific effects led to increases in RMSE compared to designs using only population-level effects with notably larger percent increases for the CPath model compared to the *de novo* BR and MERF models. Across models when only using population effects, there was no observable difference in RMSE; however, when imputing subject-specific effects, the CPath model observed increases in RMSE compared to the *de novo* models for both whole trajectories and final observations. Also, under imputation, RMSE for the MERF models was higher compared to the *de novo* BR model for both prediction types. Using known subject-fitted effects when forecasting led to the least prediction error with smaller RMSE for both *de novo* BR and MERF models compared to their population-level effects only and robust imputation designs.

Patterns for changes in AV bias for the meta-database were less consistent than those for error. When comparing within models between the population-level effects only and robust imputation designs, the CPath model observed an increase in all samplings in AVB under imputation for both whole subject trajectories and final observation forecasting while the *de novo* BR method had decreases in bias under imputation. For trajectories, there was no difference in bias within the MERF models although there was a decrease in bias in 92% of the meta-database samplings for observational forecasting.

When using fitted subject-specific effects for forecasting, the DN BR model saw decreases in AVB compared to its other two designs, meeting the 90% meta-database sampling threshold (98% for imputation and all samplings for population-level effects). This threshold also held for the MERF model comparing fitted effects to population level effects (91% of samplings) but not when comparing fitted effects to the imputed design (67.5% of samplings). Across models when using only population-level effects, although median AVB was largest in the CPath model for whole trajectories and largest in DN BR model when forecasting, neither met the 90% consistency threshold compared to AVB in the two *de novo* models. Under imputation, the CPath model had higher AVB compared to both *de novo* models for both prediction types. There was no observed difference in AVB between the *de novo* BR and MERF models under imputation for either prediction type nor was there a difference in AVB when using fitted subject-specific effects.

Synthetic Validation Datasets Results

Summary evaluation results of RMSE, MAE and AVB for the 500 60-month simulation datasets are presented for both prediction methods with average performance metrics reported in Table 90. For whole subject trajectories, percent differences in RMSE and AVB between models and the corresponding 95% confidence intervals presented in Table 91 with RMSE values visualized in Figure 83 and AVB values in Figure 84. Similar cross-design comparisons for final observation forecasts, including models with fitted subject-specific effects, are presented in Table 92 and visualized for RMSE and AVB in Figures 85 and 86 respectively. General patterns in RMSE for the meta-database were largely repeated under simulation although overall RMSE values were larger. Imputation of subject-specific effects again led to increases in RMSE compared to population-effects only designs for both types of predictions. However, of particular interest, the percent increases in RMSE for the CPath model were notably attenuated in the synthetic cohorts while percent increases were larger under simulation for the two *de novo* methods. When using only population-level effects, there was no difference in RMSE values across the various models with none of the comparisons meeting the 90% threshold. However, when using imputation of subject-specific effects, the CPath model did have higher RMSE compared to the *de novo* BR and MERF models in nearly every synthetic cohort. As before, using fitted subject-specific effects led to RMSE values significantly lower for both types of *de novo* models compared to the other subject-specific effects designs but no difference was observed in error between the *de novo* BR and MERF models when using fitted effects. This later result in particular highlights the likely dependence on width of the feature set in machine learning performance evaluation using prediction error as both *de novo* models used the same reduced covariate set, unlike the wider set for the MERF model in aim 1.

For bias under simulation, the patterns first seen in the meta-database for the *de novo* BR and MERF models were again observed. For whole trajectories, bias was lower under imputation compared to population-level only designs for the *de novo* BR model, although in only 78% of the cohorts, with no difference in bias between the MERF models. When forecasting final observations, imputation did lead to numerically lower AVB for both models although neither met the 90% threshold (74.8% of cohorts for *de novo* BR and 70.2% of cohorts for MERF). Using fitted subject specific effects led to greatly reduced AVB compared to the population-level only models for both *de novo* BR (96.6%

of cohorts) and MERF models (92% of cohorts) but did not meet the 90% cohort threshold when comparing models with subject-specific effects which were imputed to subject-specific effects fit directly (84% for DN BR and 80.2% for MERF). Most notably for the CPath model, although AVB values were still numerically larger when subject-specific effects were imputed when compared to the CPath model using only population-level effects, the increases were markedly lower under simulation compared to the meta-database, with increased bias only observed in 66% of the whole trajectory cohorts and 62.6% of the forecasting cohorts. In addition, when comparing across the various models using either population-level effects only or when imputing subject-specific effects, none of the model comparisons reached the 90% threshold which held for both whole trajectories and observational forecasting.

TABLES AND FIGURES – RESULTS

Table 3

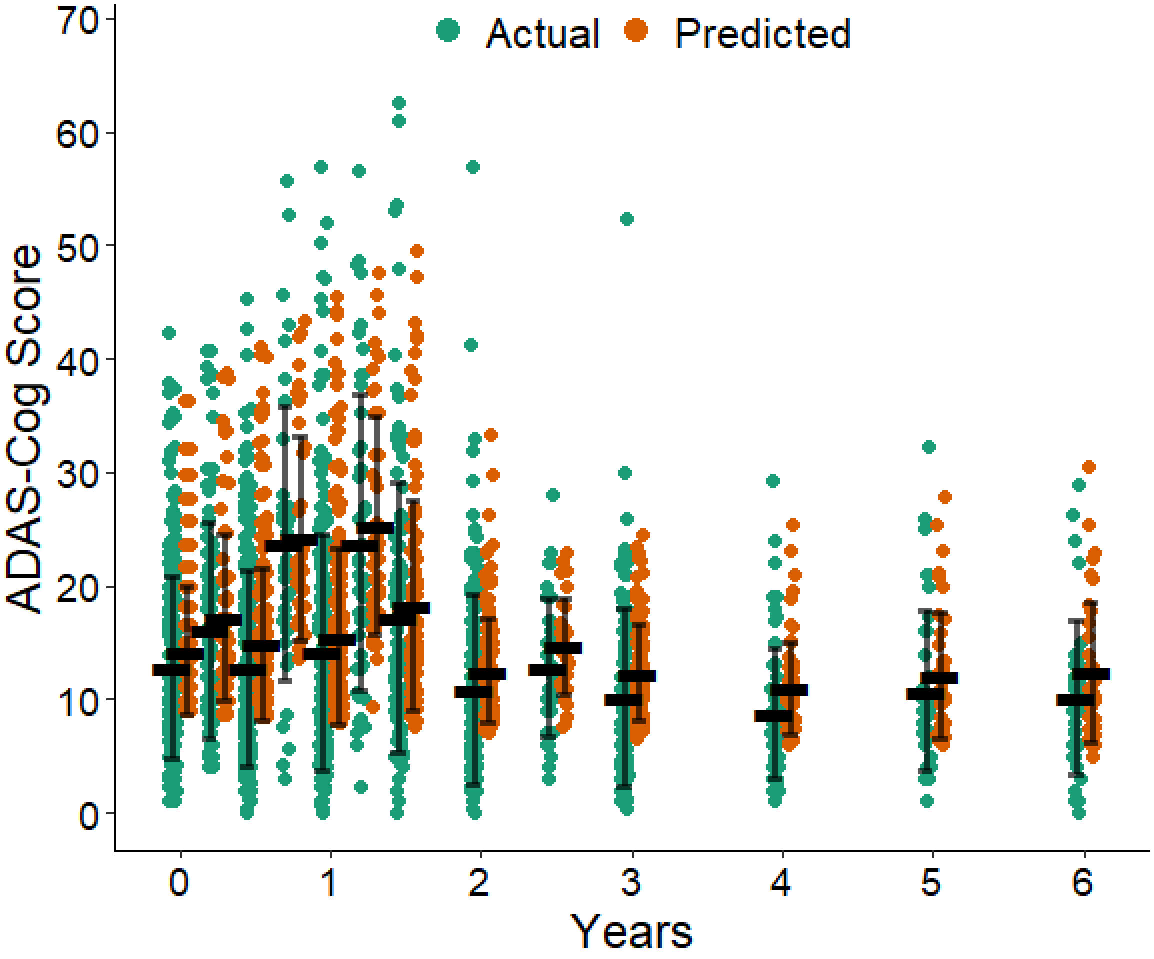
CPath Reference Model Prediction Performance – Whole Subject Trajectories of ADAS-Cog Score

Years	RMSE	MAE	Bias	AV Bias
0.0	5.222	4.220	1.481	3.668
0.25	6.054	5.035	1.159	5.004
0.5	5.851	4.753	2.088	4.163
0.75	8.957	6.950	0.434	5.439
1.0	7.109	5.407	1.383	4.422
1.25	10.142	8.260	1.579	7.643
1.5	8.228	6.416	1.025	5.390
2.0	7.082	5.448	1.601	4.620
2.5	6.294	5.024	1.832	4.251
3.0	7.541	5.760	2.119	4.727
4.0	6.076	4.828	2.228	3.951
5.0	8.072	5.982	1.328	4.756
6.0	7.817	6.034	2.272	5.017
All	6.819	5.283	1.602	4.421

Note. Regression performance metrics presented at individual time points as well as across all time points.

Figure 3

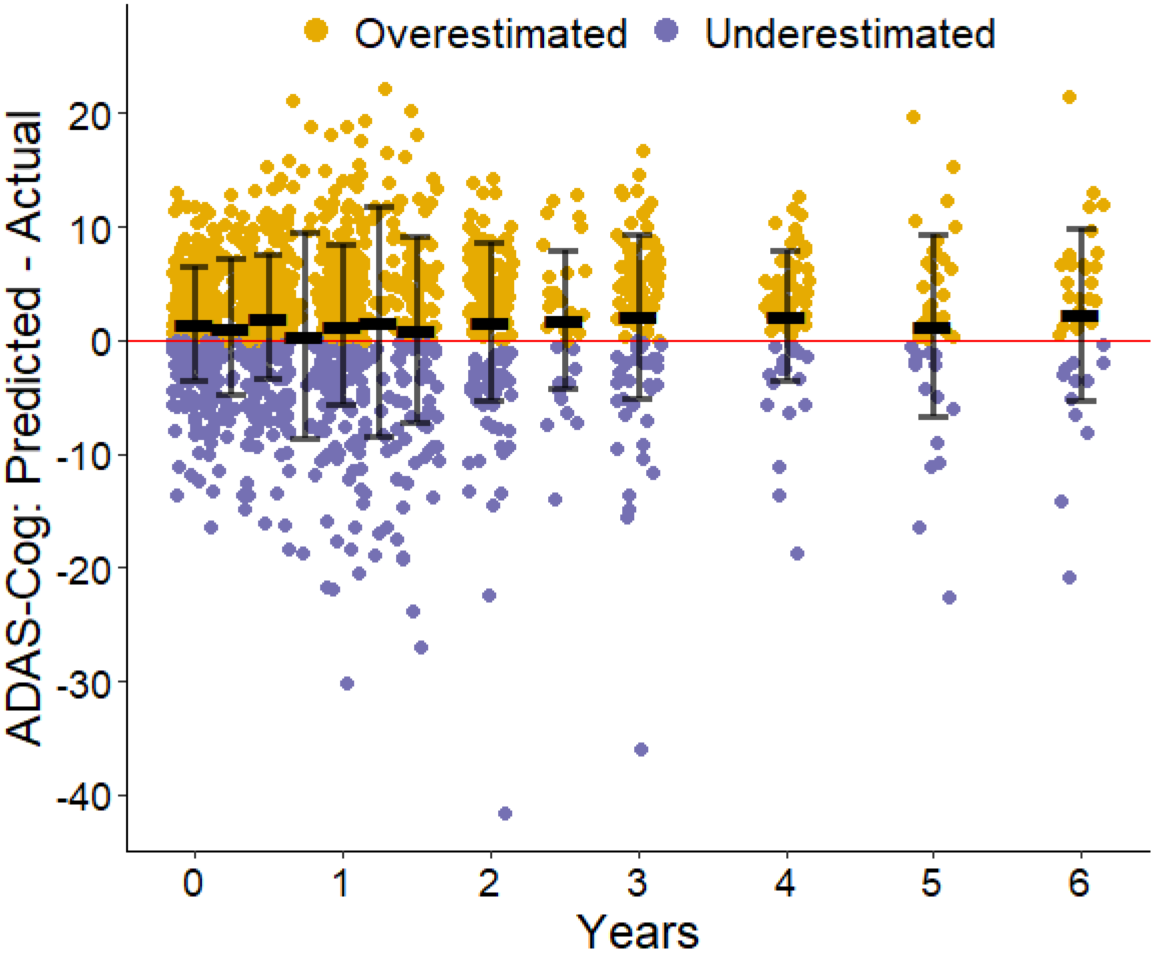
CPath Reference Model True and Predicted ADAS-Cog Scores – Whole Subject Trajectories



Note. Error bars centered at mean with standard deviation ranges.

Figure 4

CPath Reference Model Prediction Discrepancies for ADAS-Cog Scores – Whole Subject Trajectories



Note. Error bars centered at mean with standard deviation ranges.

Table 4

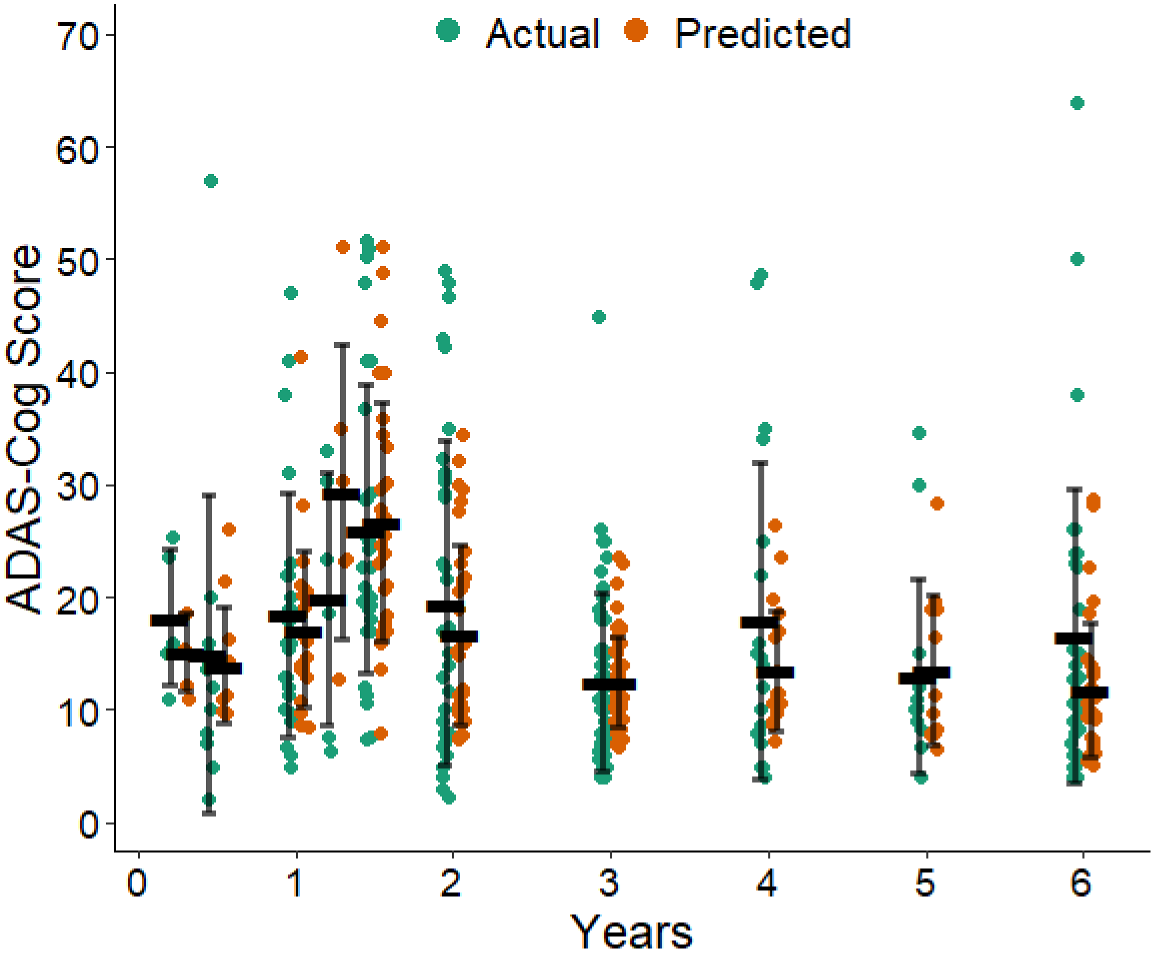
CPath Reference Model Prediction Performance – Final Observation Forecasts of ADAS-Cog Score

Years	RMSE	MAE	Bias	AV Bias
0.25	4.149	3.203	-3.054	3.733
0.5	10.578	6.994	-1.003	4.970
1.0	8.022	5.795	-1.274	4.324
1.25	15.152	9.456	9.430	3.522
1.5	10.236	7.806	0.745	6.384
2.0	9.025	6.923	-2.805	5.193
3.0	7.303	5.535	-0.010	4.783
4.0	12.497	8.489	-4.438	4.799
5.0	7.806	5.860	0.630	4.244
6.0	12.140	8.372	-4.779	5.744
All	9.720	6.859	-1.360	4.799

Note. Regression performance metrics presented at individual time points as well as across all time points.

Figure 5

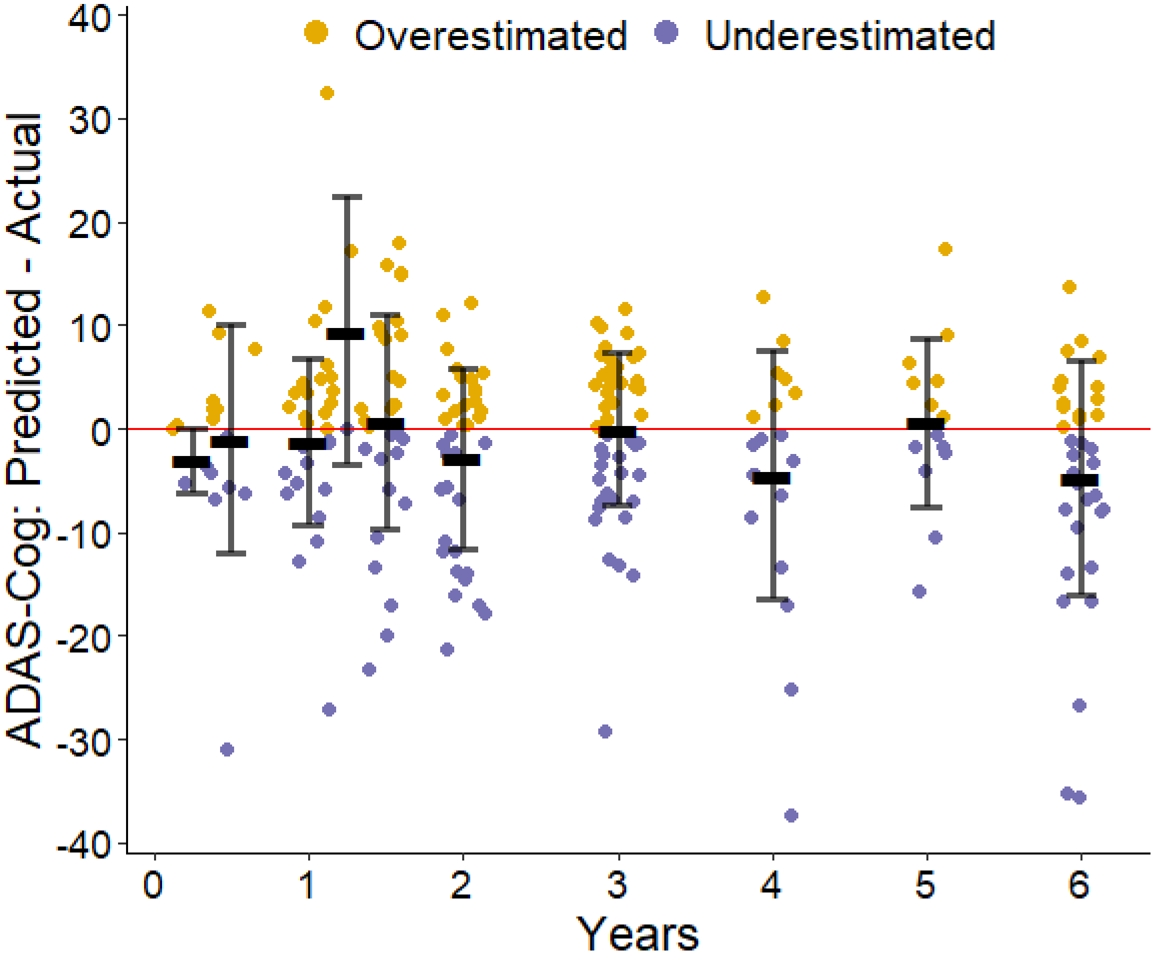
CPath Reference Model True and Predicted ADAS-Cog Scores – Final Observation Forecasts



Note. Error bars centered at mean with standard deviation ranges.

Figure 6

CPath Reference Model Prediction Discrepancies for ADAS-Cog Scores – Final Observation Forecasts



Note. Error bars centered at mean with standard deviation ranges.

Table 5

Logistic Reference Model Prediction Performance – Whole Subject Trajectories of CDR-Based Impairment

Years	Accuracy	Precision	Recall	AUC
0.0	0.747	0.984	0.705	0.879
0.5	0.747	0.973	0.711	0.866
1.0	0.734	0.964	0.704	0.861
1.5	0.750	0.976	0.757	0.753
2.0	0.657	0.919	0.564	0.827
2.5	0.586	1.000	0.586	1.000
3.0	0.581	0.912	0.484	0.767
4.0	0.706	0.778	0.350	0.695
5.0	0.606	0.750	0.353	0.669
6.0	0.594	0.571	0.286	0.639
All	0.711	0.958	0.661	0.841

Note. Classification performance metrics presented at individual time points as well as across all time points.

Figure 7

Logistic Reference Model True and Predicted CDR-Based Impairment Counts – Whole Subject Trajectories

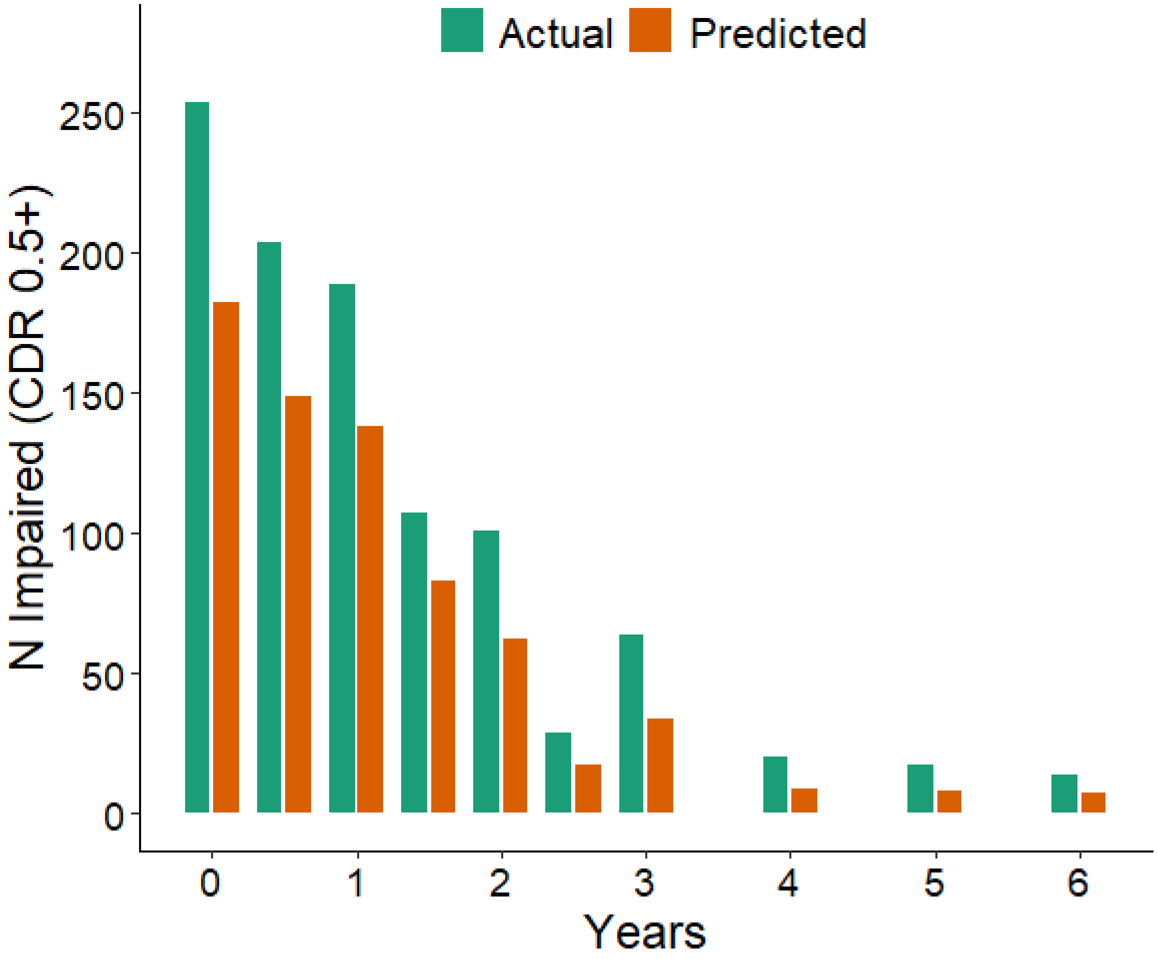
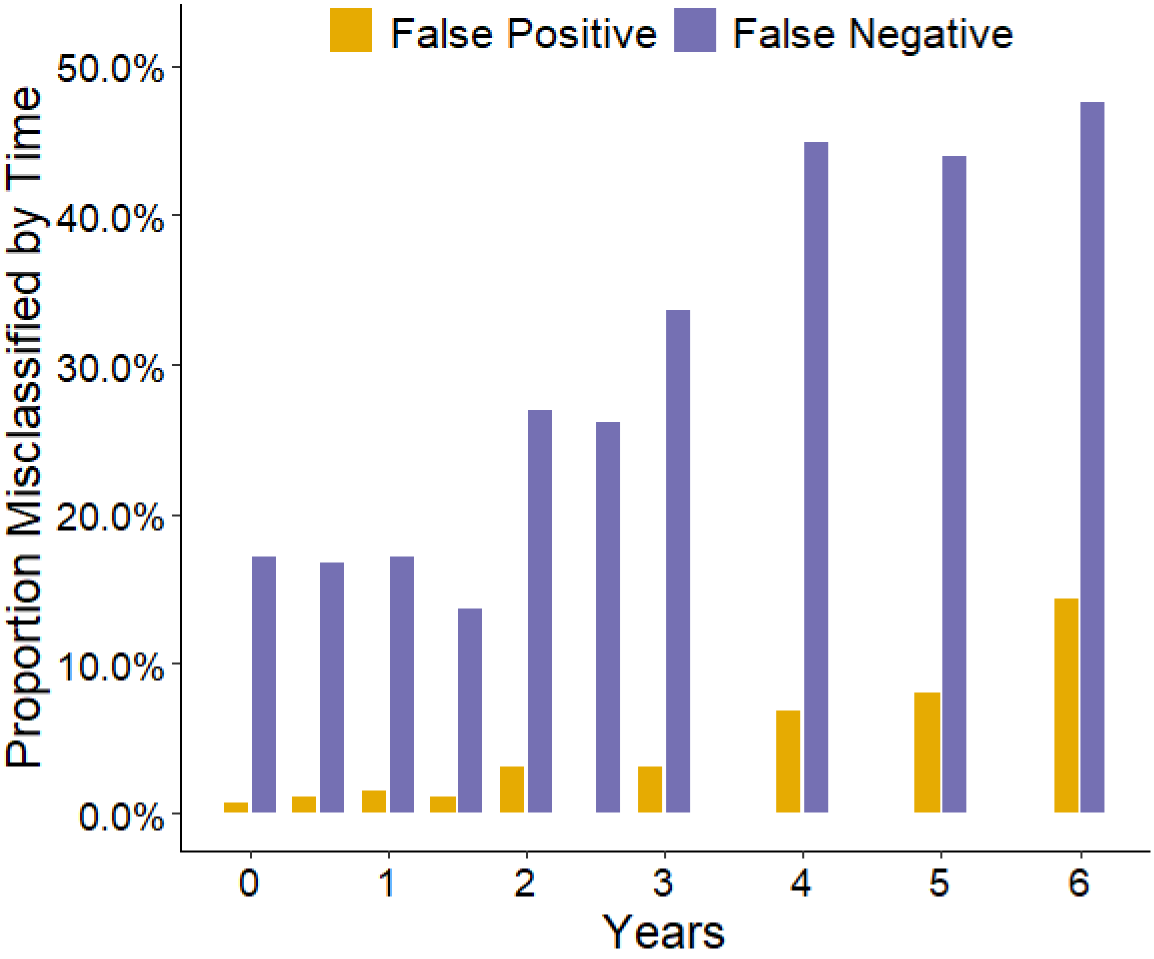


Figure 8

Logistic Reference Model Misclassification Rates for CDR-Based Impairment – Whole Subject Trajectories



Note. False positive and false negative rates marginalized by time to be relative to total counts at that timepoint.

Table 6

Logistic Reference Model Prediction Performance – Final Observation Forecasts of CDR-Based Impairment

Years	Accuracy	Precision	Recall	AUC
0.5	0.571	0.889	0.615	0.615
1.0	0.774	1.000	0.774	1.000
1.5	0.967	1.000	0.967	1.000
2.0	0.806	0.960	0.800	0.867
3.0	0.587	0.957	0.550	0.662
4.0	0.667	1.000	0.500	0.700
5.0	0.833	1.000	0.714	0.829
6.0	0.645	1.000	0.421	0.794
All	0.730	0.977	0.694	0.825

Note. Classification performance metrics presented at individual time points as well as across all time points.

Figure 9

Logistic Reference Model True and Predicted CDR-Based Impairment Counts – Final Observation Forecasts

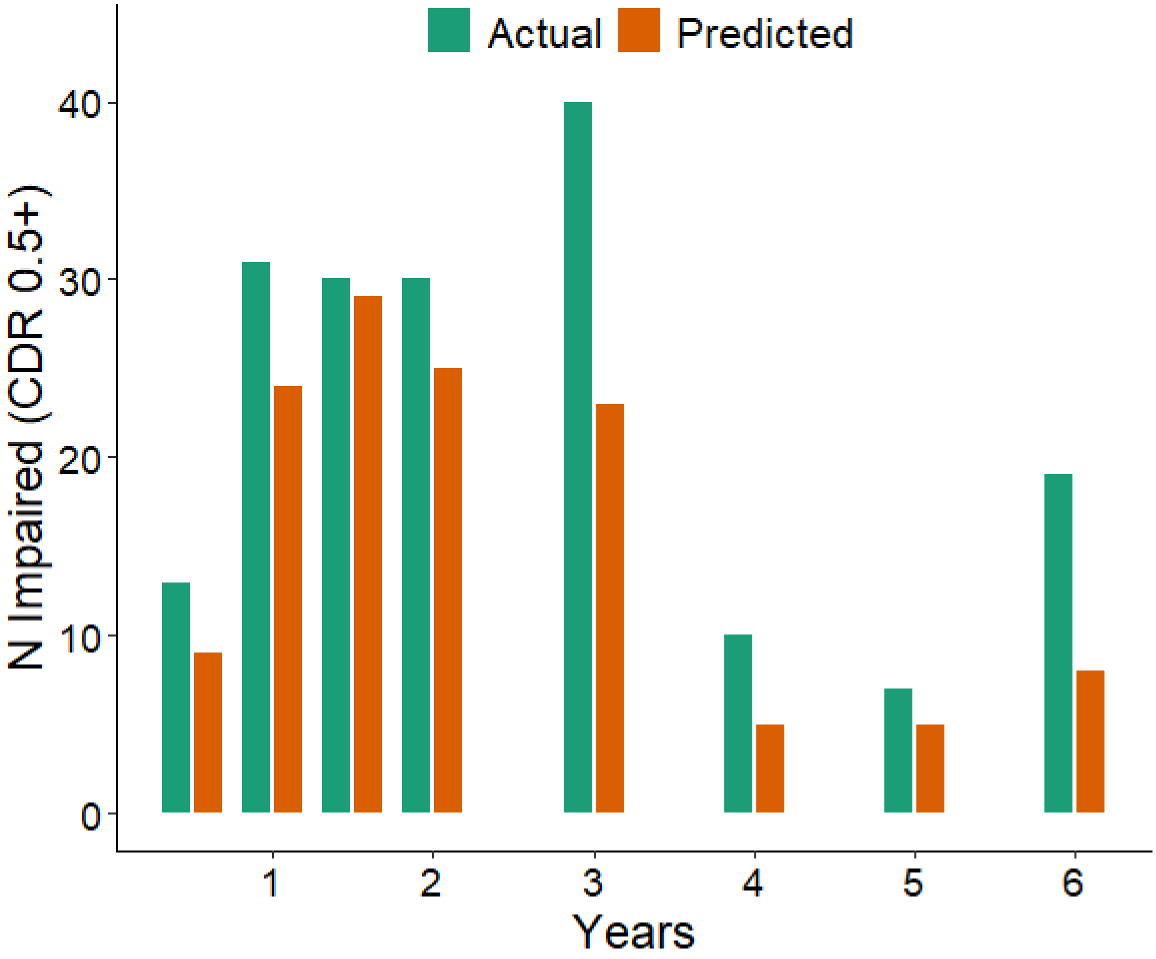
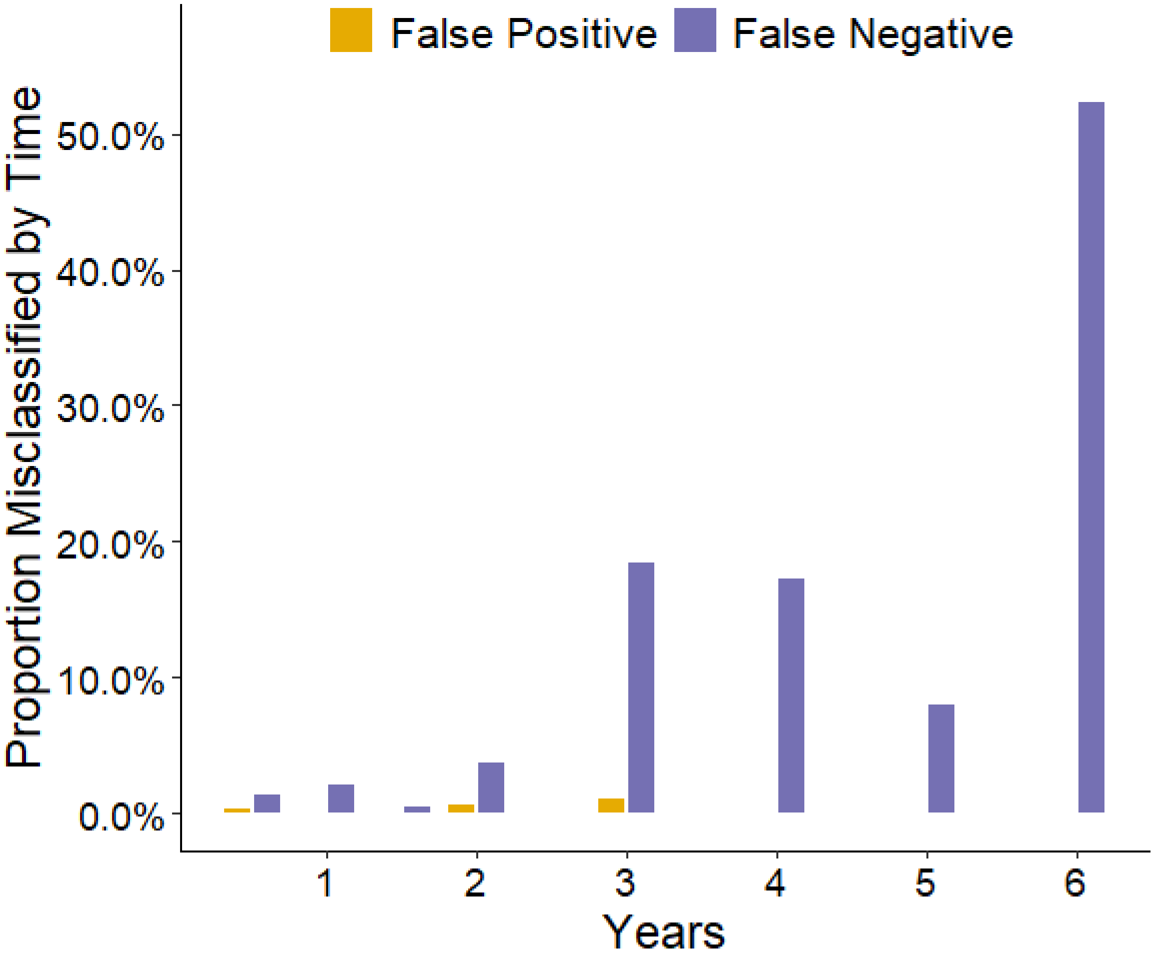


Figure 10

Logistic Reference Model Misclassification Rates for CDR-Based Impairment – Final Observation Forecasts



Note. False positive and false negative rates marginalized by time to be relative to total counts at that timepoint.

Table 7

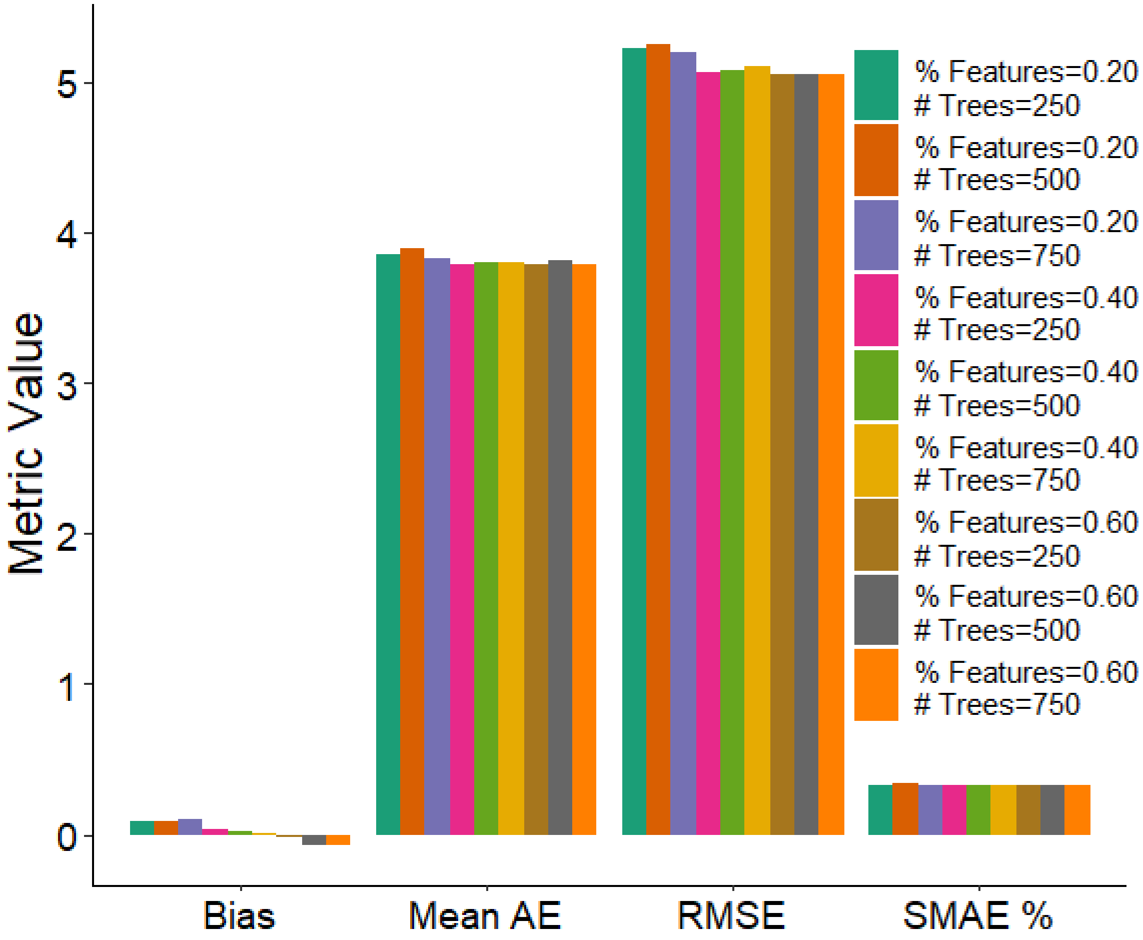
Ensemble Methods Performance Summary – ADAS-Cog Score

Type of prediction	Performance metric	CPath reference	MERF	Single GLMM	Bagged GLMM	Boosted trees
Whole subject trajectories						
	RMSE	6.819	4.928	5.684	5.542	4.943
	Mean AE	5.283	3.792	4.427	4.323	3.858
	Bias	1.602	0.066	0.427	0.423	0.107
	AV bias	4.421	3.162	3.673	3.557	3.227
Final observation forecasts						
	RMSE	9.720	4.721	4.528	5.142	5.233
	Mean AE	6.859	3.588	3.357	3.842	3.971
	Bias	-1.360	-1.151	-0.593	-0.208	-1.322
	AV bias	4.799	2.820	2.573	2.888	3.158

Note. Regression performance metrics summarized across all time points with CPath model shown for reference.

Figure 11

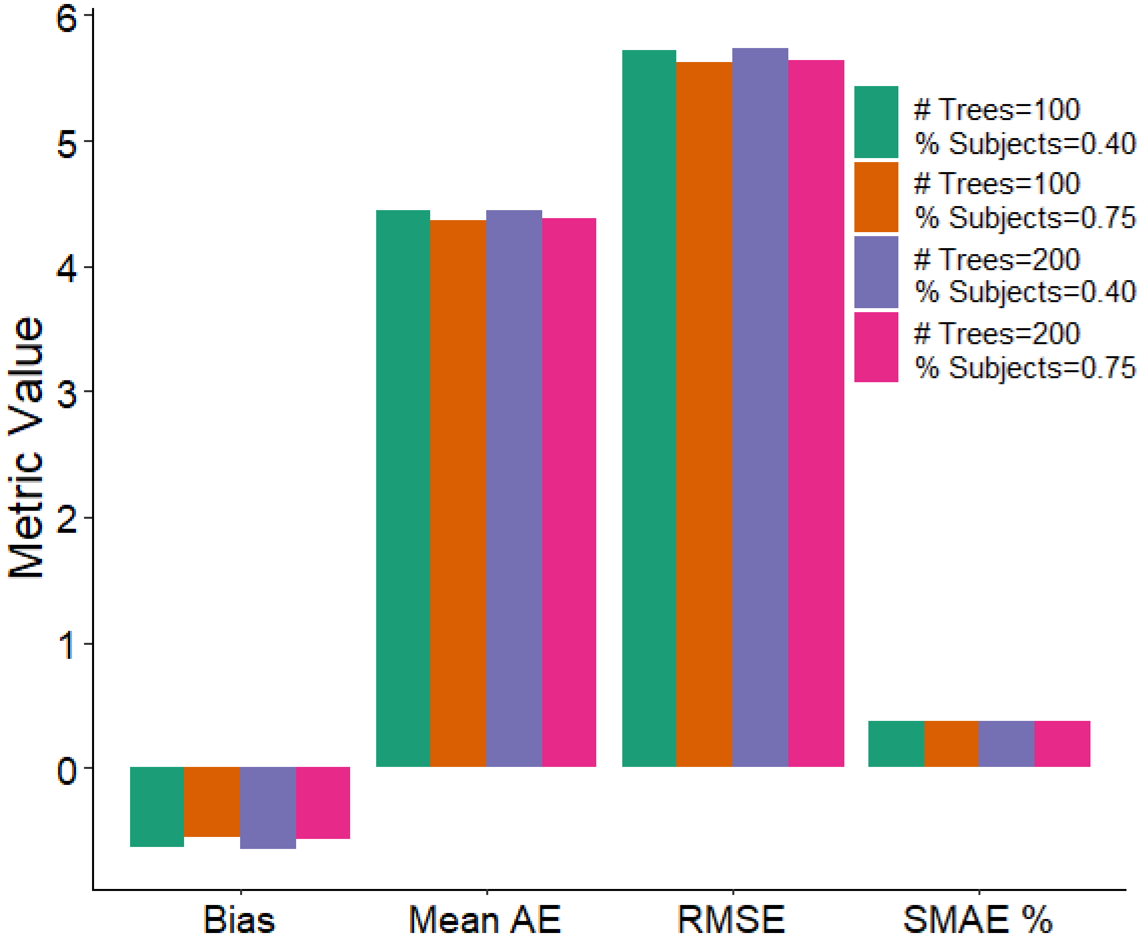
MERF Model Hyperparameter Tunings – ADAS-Cog Score



Note. Hyperparameter tunings provided from cross-fold validations on whole subject trajectories.

Figure 12

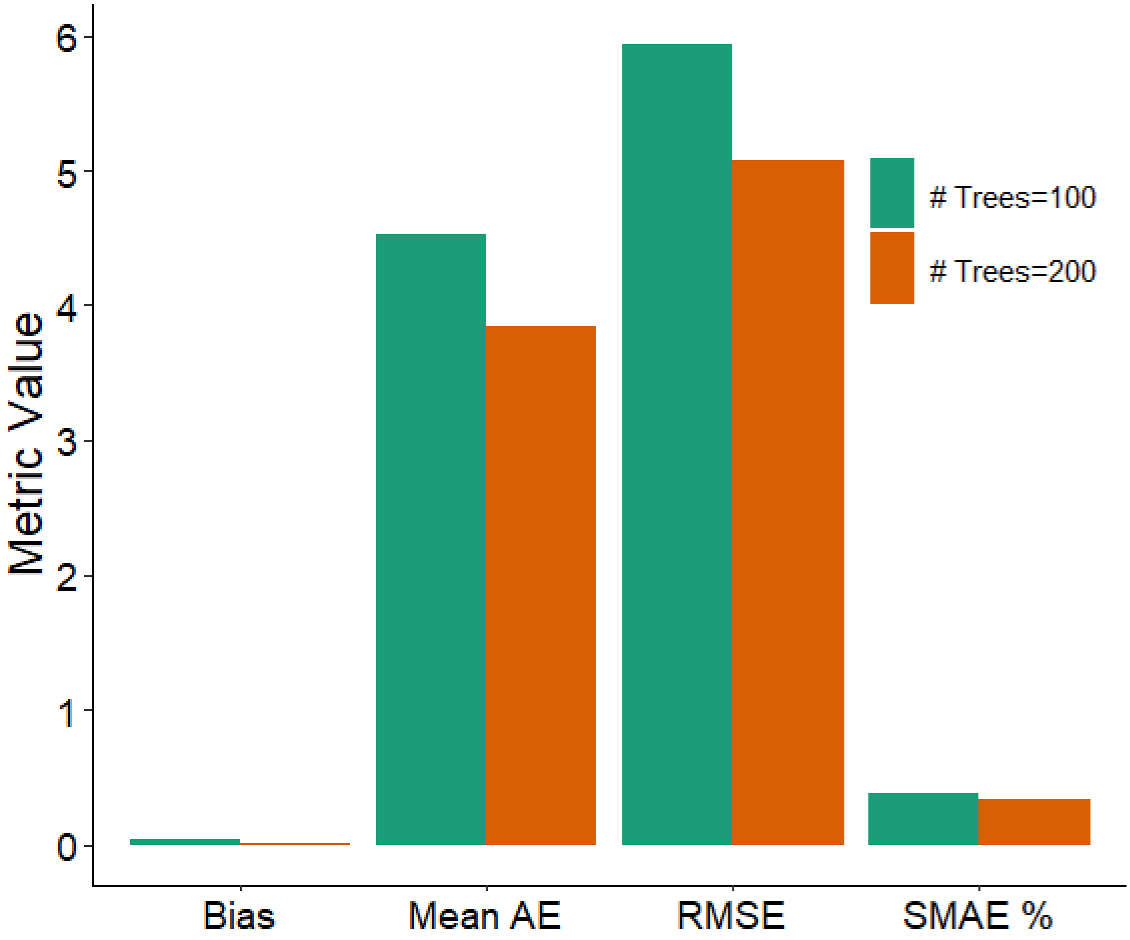
Bagged GLMM Trees Model Hyperparameter Tunings – ADAS-Cog Score



Note. Hyperparameter tunings provided from cross-fold validations on whole subject trajectories.

Figure 13

Boosted Trees Model Hyperparameter Tunings – ADAS-Cog Score



Note. Hyperparameter tunings provided from cross-fold validations on whole subject trajectories.

Table 8

Top Hyperparameters for Ensemble Methods – ADAS-Cog Score

Ensemble method	Performance metric	Best value	Hyperparameter set
MERF model			
	RMSE	5.044	% Features=0.60 ; # Trees=250
	Mean AE	3.791	% Features=0.60 ; # Trees=750
	SMAE %	.330	% Features=0.40 ; # Trees=250
	Bias	-0.012	% Features=0.40 ; # Trees=750
Bagged GLMM trees model			
	RMSE	5.626	# Trees=100 ; % Subjects=0.75
	Mean AE	4.363	# Trees=100 ; % Subjects=0.75
	SMAE %	0.365	# Trees=100 ; % Subjects=0.75
	Bias	-0.548	# Trees=100 ; % Subjects=0.75
Boosted trees model			
	RMSE	5.074	# Trees=200
	Mean AE	3.835	# Trees=200
	SMAE %	0.337	# Trees=200
	Bias	-0.013	# Trees=200

Note. For each ensemble method, the best performing hyperparameter set for each regression metric is displayed along with the corresponding value.

Table 9

MERF Model Prediction Performance – Whole Subject Trajectories of ADAS-Cog Score

Years	RMSE	MAE	Bias	AV Bias
0.0	4.749	3.628	0.096	2.885
0.25	5.381	4.211	-1.007	3.407
0.5	4.486	3.439	0.491	2.916
0.75	6.605	5.080	-2.476	3.759
1.0	4.989	3.820	-0.127	3.226
1.25	6.968	5.668	-1.849	4.412
1.5	5.337	3.902	-0.564	3.059
2.0	4.830	3.829	0.641	3.490
2.5	4.907	3.869	-2.046	2.766
3.0	4.856	3.954	0.527	3.668
4.0	3.880	3.141	1.410	3.004
5.0	3.561	2.767	1.717	2.133
6.0	4.541	3.665	2.090	2.968
All	4.928	3.792	0.066	3.162

Note. Regression performance metrics presented at individual time points as well as across all time points.

Table 10

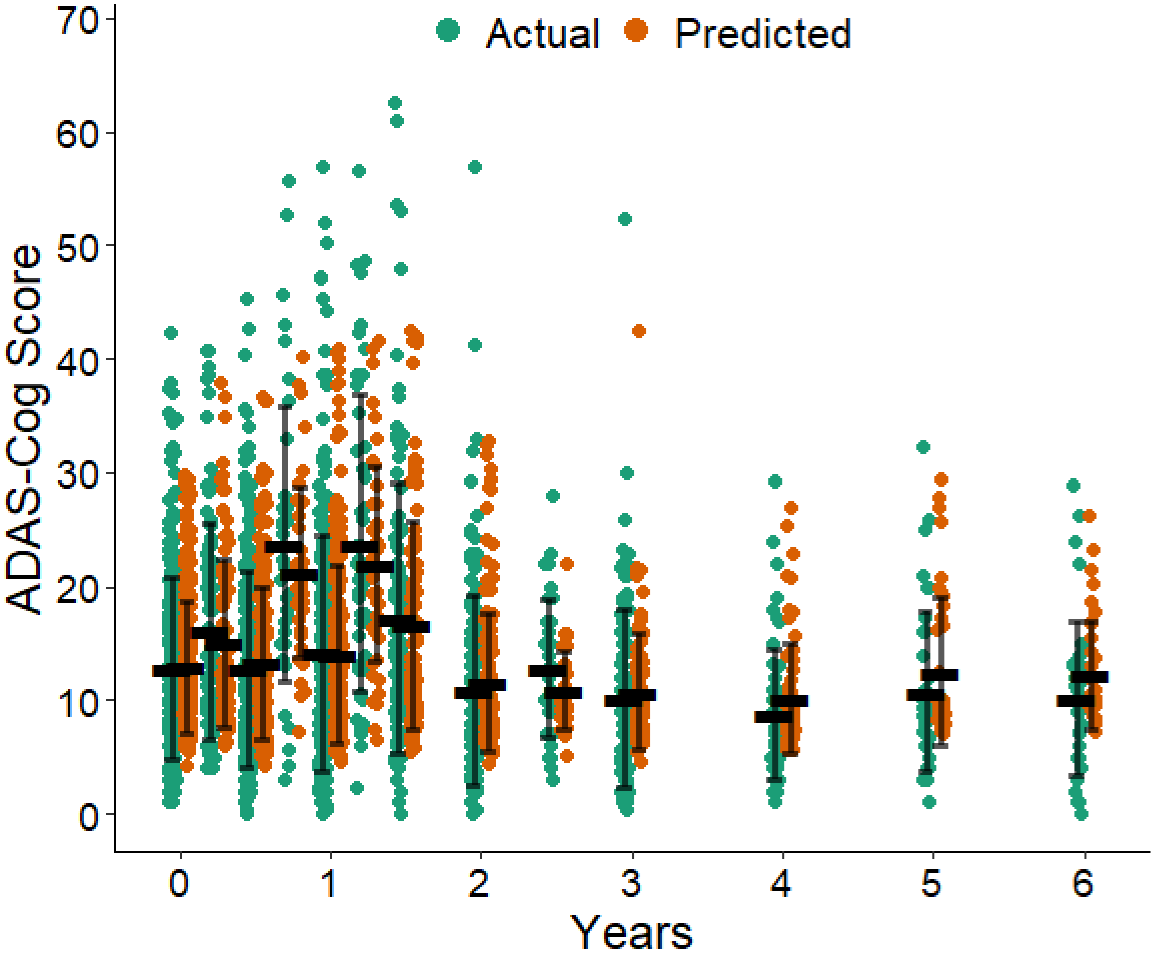
Evaluation of MERF Model Relative to CPath Reference – Whole Subject Trajectories of ADAS-Cog Score

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
RMSE	4.928	6.819	-1.891	-27.73%	[-2.112, -1.658]
Mean AE	3.792	5.283	-1.491	-28.23%	[-1.651, -1.321]
Bias	0.066	1.602	-1.536	-95.90%	[-1.797, -1.289]
AV bias	3.162	4.421	-1.258	-28.46%	[-1.449, -1.036]

Note. Values for change in metric and percent difference relative to CPath reference model; bootstrap confidence interval corresponds to change in metric.

Figure 14

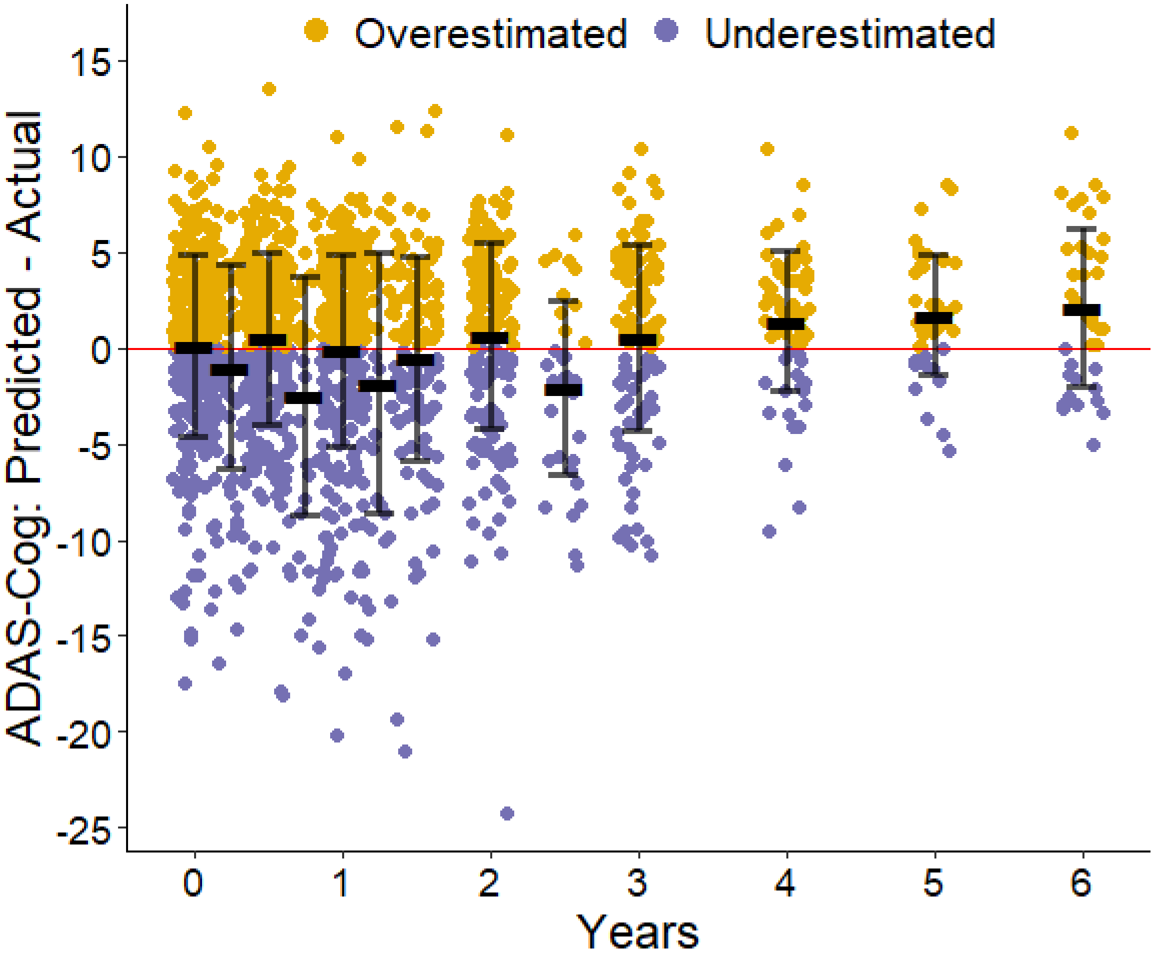
MERF Model True and Predicted ADAS-Cog Scores – Whole Subject Trajectories



Note. Error bars centered at mean with standard deviation ranges.

Figure 15

MERF Model Prediction Discrepancies for ADAS-Cog Scores – Whole Subject Trajectories



Note. Error bars centered at mean with standard deviation ranges.

Table 11

MERF Model Prediction Performance – Final Observation Forecasts of ADAS-Cog Score

Years	RMSE	MAE	Bias	AV Bias
0.25	3.407	2.728	-0.008	2.545
0.5	4.692	3.171	-0.880	2.436
1.0	3.982	3.262	-0.703	2.736
1.25	6.023	5.084	-0.320	6.509
1.5	6.031	4.877	-0.681	4.230
2.0	4.995	3.670	-1.505	2.269
3.0	4.108	3.044	-0.892	2.340
4.0	5.275	3.858	-1.793	2.876
5.0	3.089	2.388	-1.556	2.268
6.0	4.660	3.821	-1.818	3.043
All	4.721	3.588	-1.151	2.820

Note. Regression performance metrics presented at individual time points as well as across all time points.

Table 12

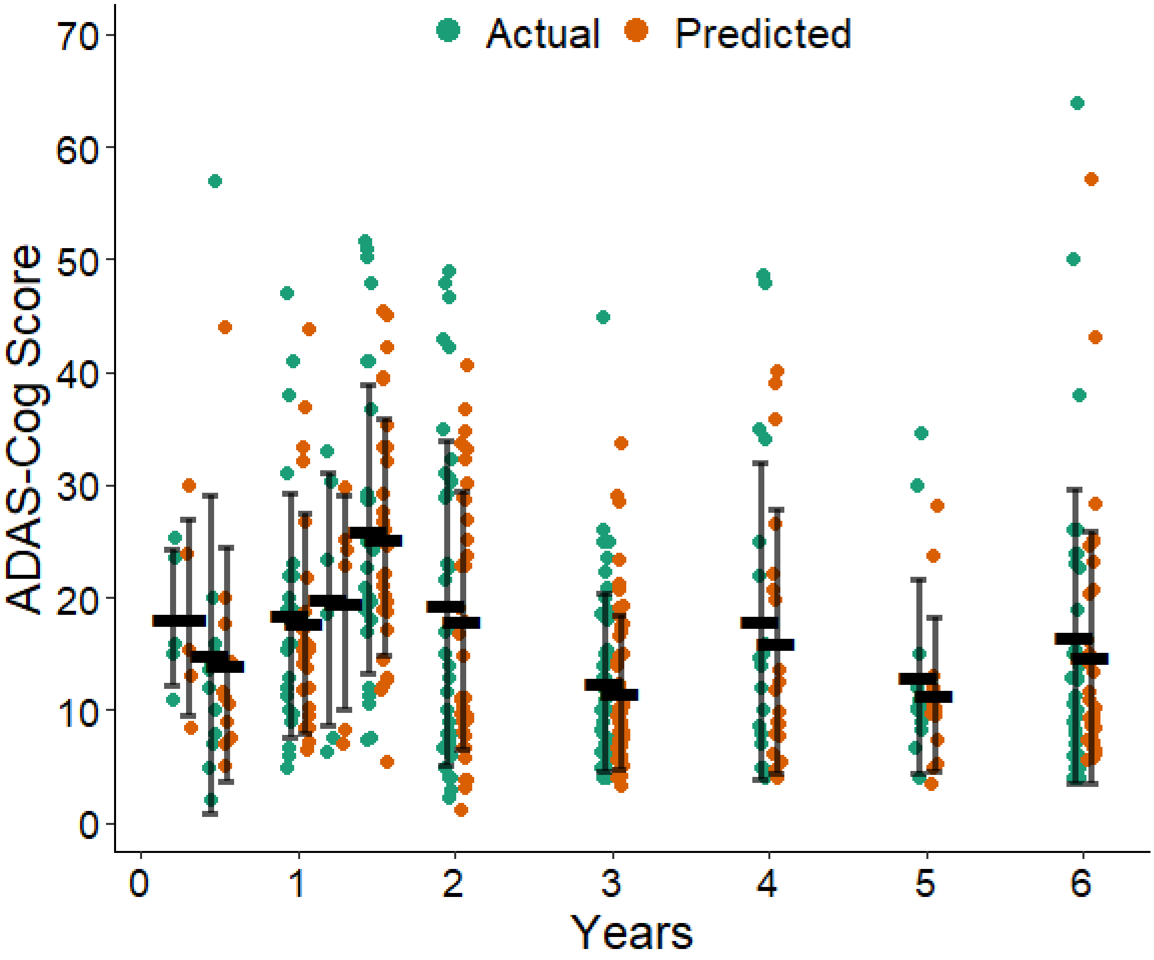
Evaluation of MERF Model Relative to CPath Reference – Final Observation Forecasts of ADAS-Cog Score

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
RMSE	4.721	9.720	-4.999	-51.43%	[-5.490, -4.532]
Mean AE	3.588	6.859	-3.271	-47.69%	[-3.635, -2.899]
Bias	-1.151	-1.360	0.209	-15.37%	[-0.350, 0.796]
AV bias	2.820	4.799	-1.979	-41.24%	[-2.393, -1.649]

Note. Values for change in metric and percent difference relative to CPath reference model; bootstrap confidence interval corresponds to change in metric.

Figure 16

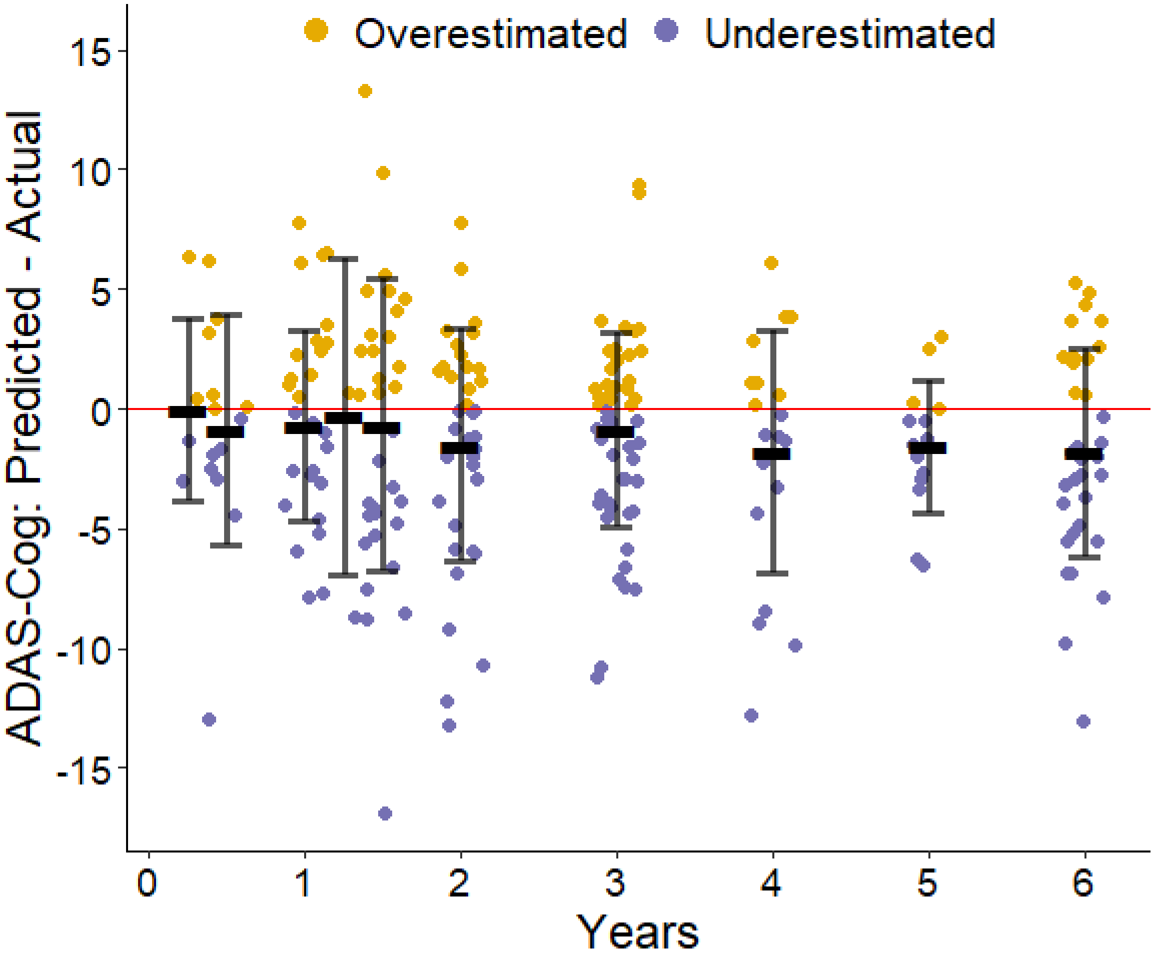
MERF Model True and Predicted ADAS-Cog Scores – Final Observation Forecasts



Note. Error bars centered at mean with standard deviation ranges.

Figure 17

MERF Model Prediction Discrepancies for ADAS-Cog Scores – Final Observation Forecasts



Note. Error bars centered at mean with standard deviation ranges.

Table 13

Single GLMM Tree Model Prediction Performance – Whole Subject Trajectories of ADAS-Cog Score

Years	RMSE	MAE	Bias	AV Bias
0.0	5.039	3.770	-0.165	2.811
0.25	5.949	4.675	-1.653	3.932
0.5	4.987	3.882	0.460	3.462
0.75	7.579	5.878	-3.290	4.255
1.0	5.609	4.438	0.181	3.461
1.25	8.007	6.232	-2.928	4.956
1.5	6.377	4.710	-0.626	3.756
2.0	5.528	4.592	1.742	4.205
2.5	5.088	4.082	-1.775	3.975
3.0	5.919	4.910	2.139	4.576
4.0	5.550	4.672	3.687	4.354
5.0	5.844	4.704	4.394	3.830
6.0	6.857	5.332	4.829	4.522
All	5.684	4.427	0.427	3.673

Note. Regression performance metrics presented at individual time points as well as across all time points.

Table 14

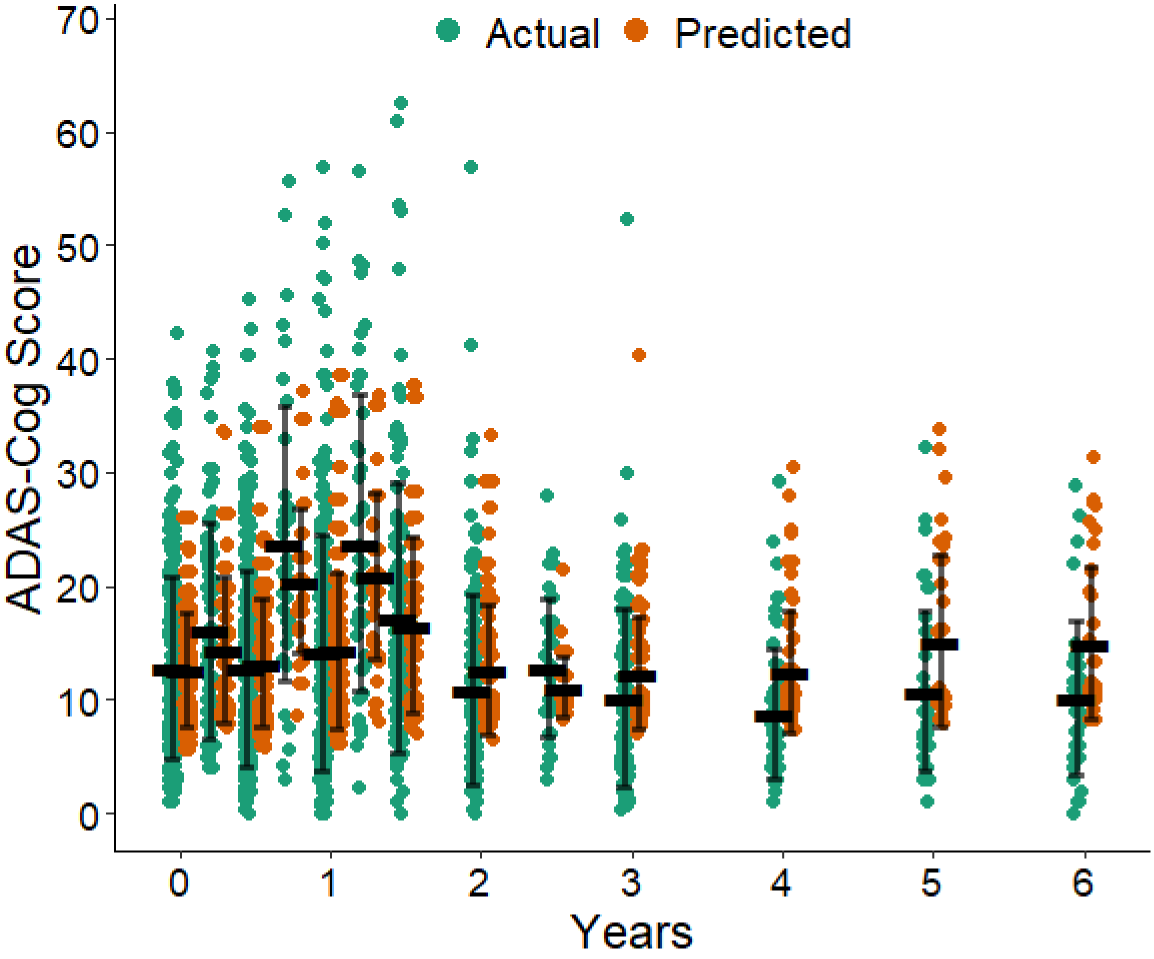
Evaluation of Single GLMM Tree Model Relative to CPath Reference – Whole Subject Trajectories of ADAS-Cog Score

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
RMSE	5.684	6.819	-1.135	-16.64%	[-1.378, -0.890]
Mean AE	4.427	5.283	-0.856	-16.20%	[-1.031, -0.681]
Bias	0.427	1.602	-1.175	-73.35%	[-1.468, -0.906]
AV bias	3.673	4.421	-0.748	-16.92%	[-0.992, -0.532]

Note. Values for change in metric and percent difference relative to CPath reference model; bootstrap confidence interval corresponds to change in metric.

Figure 18

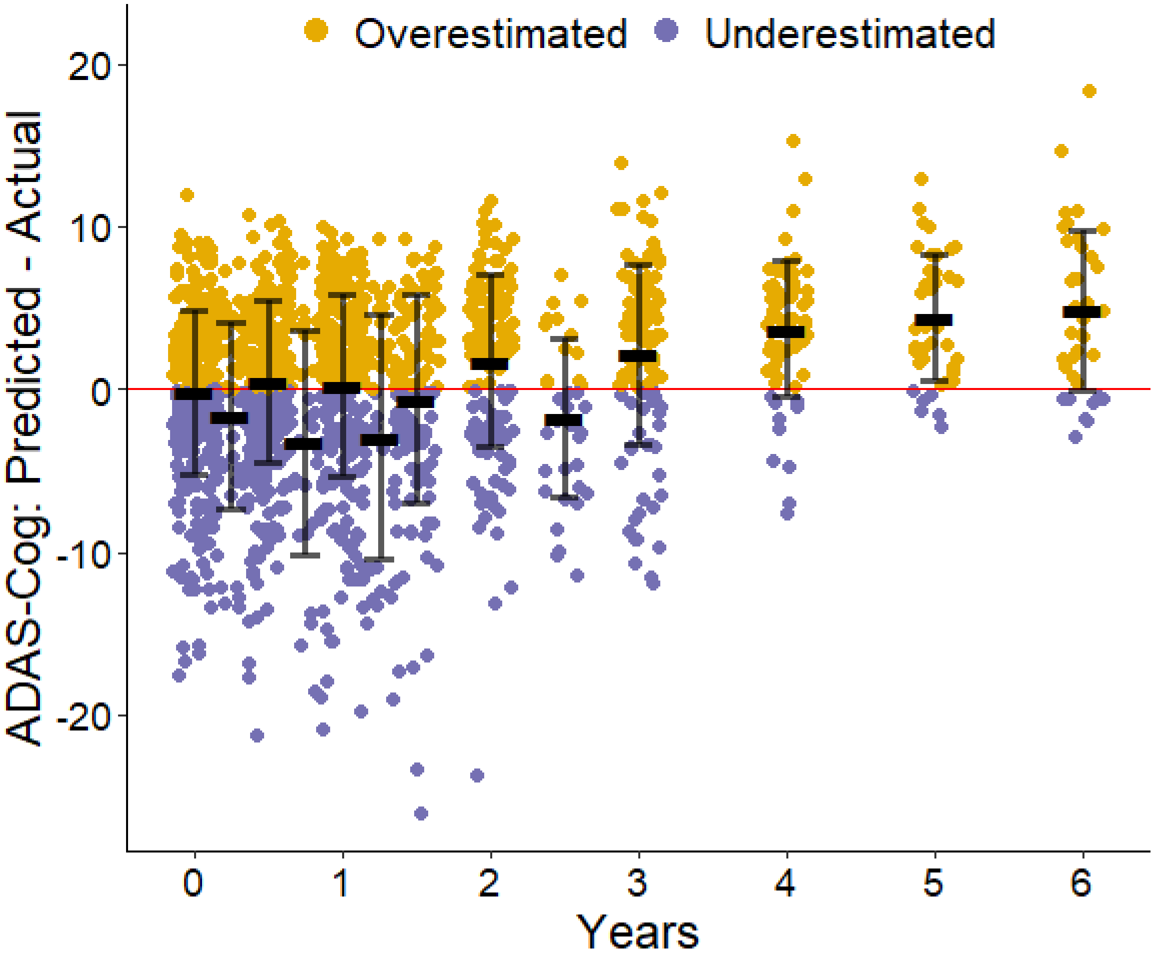
Single GLMM Tree Model True and Predicted ADAS-Cog Scores – Whole Subject Trajectories



Note. Error bars centered at mean with standard deviation ranges.

Figure 19

Single GLMM Tree Model Prediction Discrepancies for ADAS-Cog Scores – Whole Subject Trajectories



Note. Error bars centered at mean with standard deviation ranges.

Table 15

Single GLMM Tree Model Prediction Performance – Final Observation Forecasts of ADAS-Cog Score

Years	RMSE	MAE	Bias	AV Bias
0.25	3.768	3.336	-1.000	3.704
0.5	4.847	3.153	-0.261	2.158
1.0	3.662	2.828	-0.554	1.908
1.25	5.609	4.595	0.661	4.452
1.5	6.334	4.789	-1.004	3.253
2.0	5.290	3.837	-1.073	2.591
3.0	3.846	2.851	-0.506	2.166
4.0	4.181	3.148	-0.385	2.759
5.0	2.630	2.142	-0.697	1.962
6.0	3.842	3.253	-0.240	2.908
All	4.528	3.357	-0.593	2.573

Note. Regression performance metrics presented at individual time points as well as across all time points.

Table 16

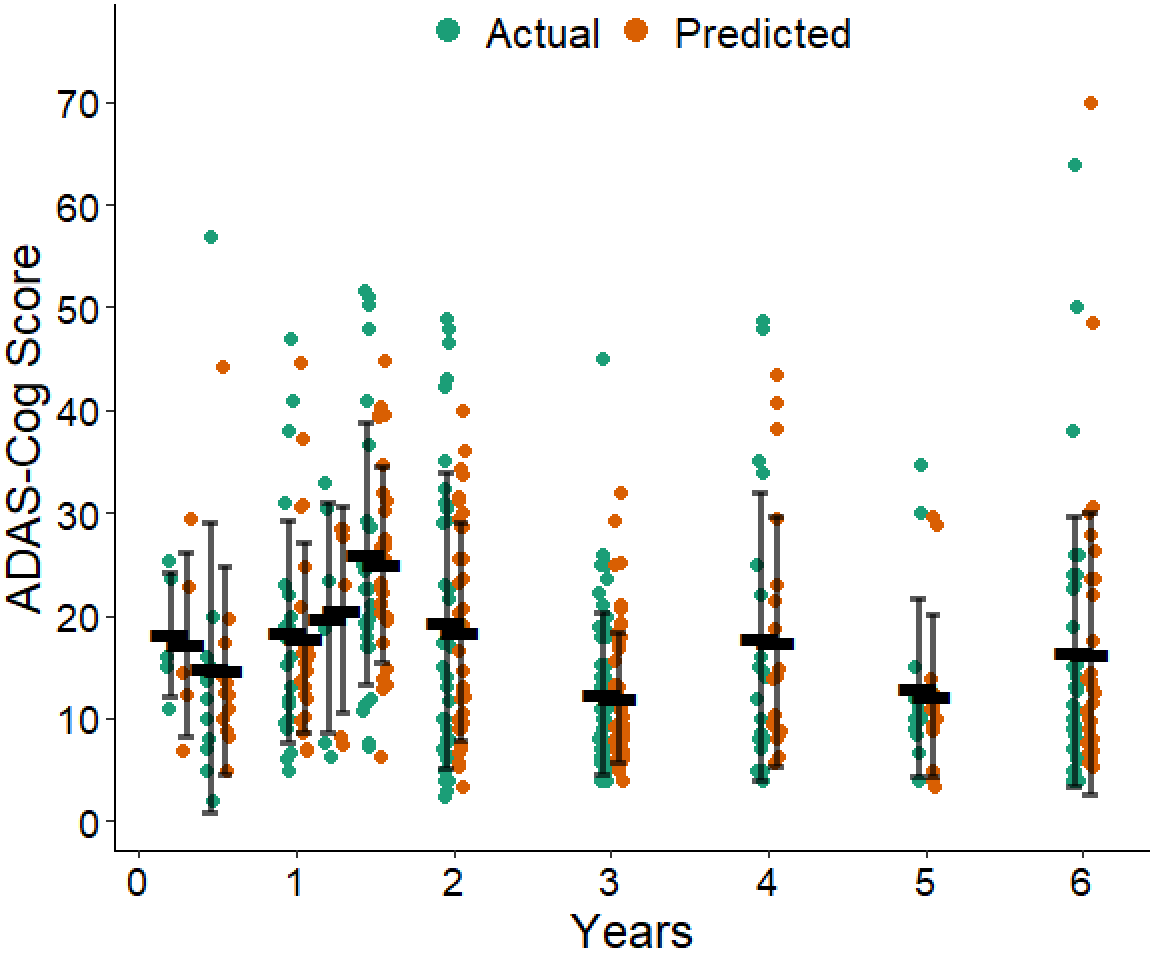
Evaluation of Single GLMM Tree Model Relative to CPath Reference – Final Observation Forecasts of ADAS-Cog Score

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
RMSE	4.528	9.720	-5.192	-53.42%	[-5.767, -4.665]
Mean AE	3.357	6.859	-3.501	-51.05%	[-3.885, -3.105]
Bias	-0.593	-1.360	0.768	-56.43%	[0.216, 1.331]
AV bias	2.573	4.799	-2.226	-46.39%	[-2.464, -1.839]

Note. Values for change in metric and percent difference relative to CPath reference model; bootstrap confidence interval corresponds to change in metric.

Figure 20

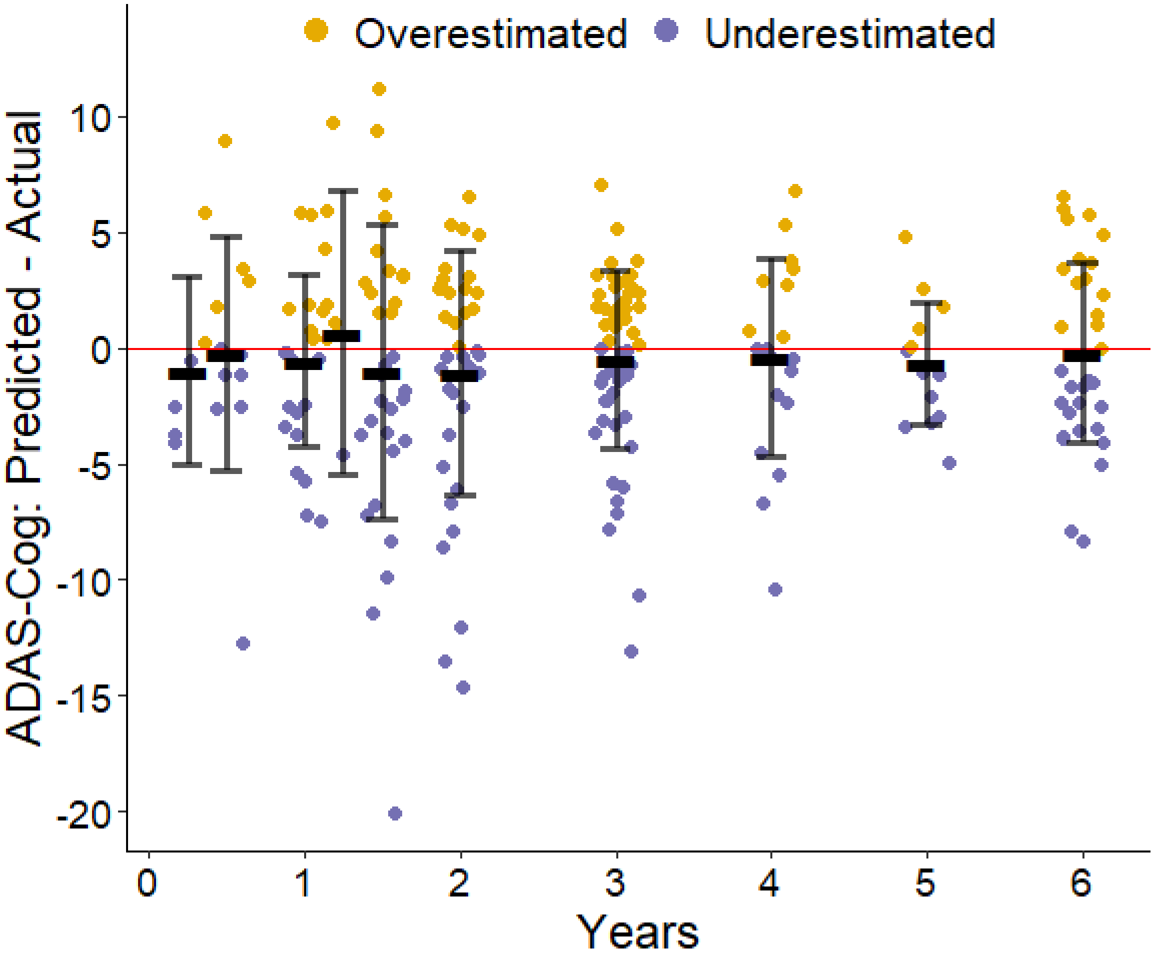
Single GLMM Tree Model True and Predicted ADAS-Cog Scores – Final Observation Forecasts



Note. Error bars centered at mean with standard deviation ranges.

Figure 21

Single GLMM Tree Model Prediction Discrepancies for ADAS-Cog Scores – Final Observation Forecasts



Note. Error bars centered at mean with standard deviation ranges.

Table 17

Bagged GLMM Trees Model Prediction Performance – Whole Subject Trajectories of ADAS-Cog Score

Years	RMSE	MAE	Bias	AV Bias
0.0	4.995	3.737	-0.145	2.667
0.25	5.913	4.658	-1.638	3.783
0.5	4.925	3.833	0.489	3.260
0.75	7.531	5.818	-3.175	4.244
1.0	5.495	4.351	0.172	3.494
1.25	7.711	6.105	-2.724	5.057
1.5	6.061	4.589	-0.514	3.639
2.0	5.426	4.480	1.670	4.030
2.5	4.840	3.816	-1.634	3.817
3.0	5.690	4.689	2.151	4.597
4.0	5.366	4.493	3.526	4.232
5.0	5.380	4.264	4.012	3.246
6.0	6.424	5.012	4.426	3.560
All	5.542	4.323	0.423	3.557

Note. Regression performance metrics presented at individual time points as well as across all time points.

Table 18

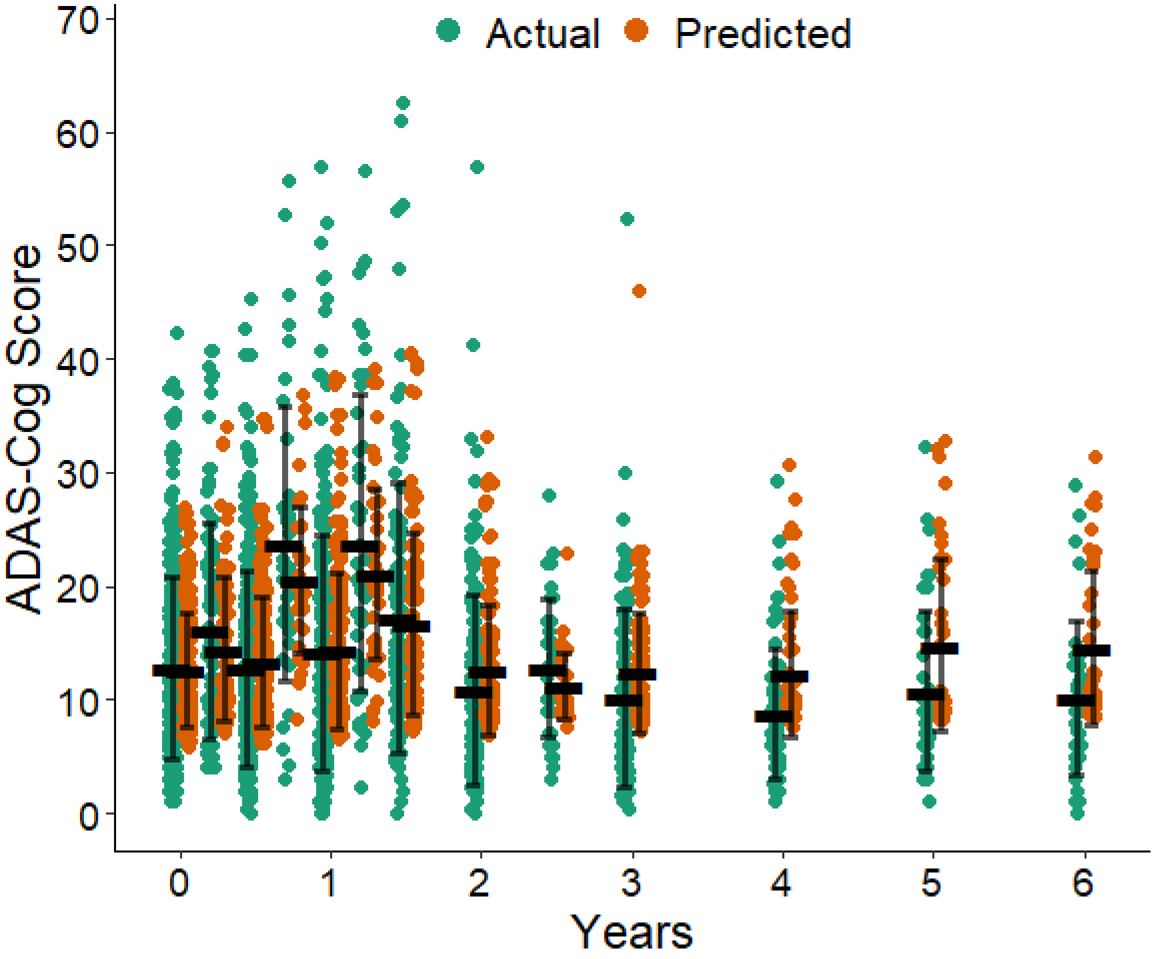
Evaluation of Bagged GLMM Trees Model Relative to CPath Reference – Whole Subject Trajectories of ADAS-Cog Score

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
RMSE	5.542	6.819	-1.277	-18.73%	[-1.503, -1.041]
Mean AE	4.323	5.283	-0.960	-18.18%	[-1.131, -0.786]
Bias	0.423	1.602	-1.179	-73.61%	[-1.487, -0.897]
AV bias	3.557	4.421	-0.864	-19.55%	[-1.079, -0.604]

Note. Values for change in metric and percent difference relative to CPath reference model; bootstrap confidence interval corresponds to change in metric.

Figure 22

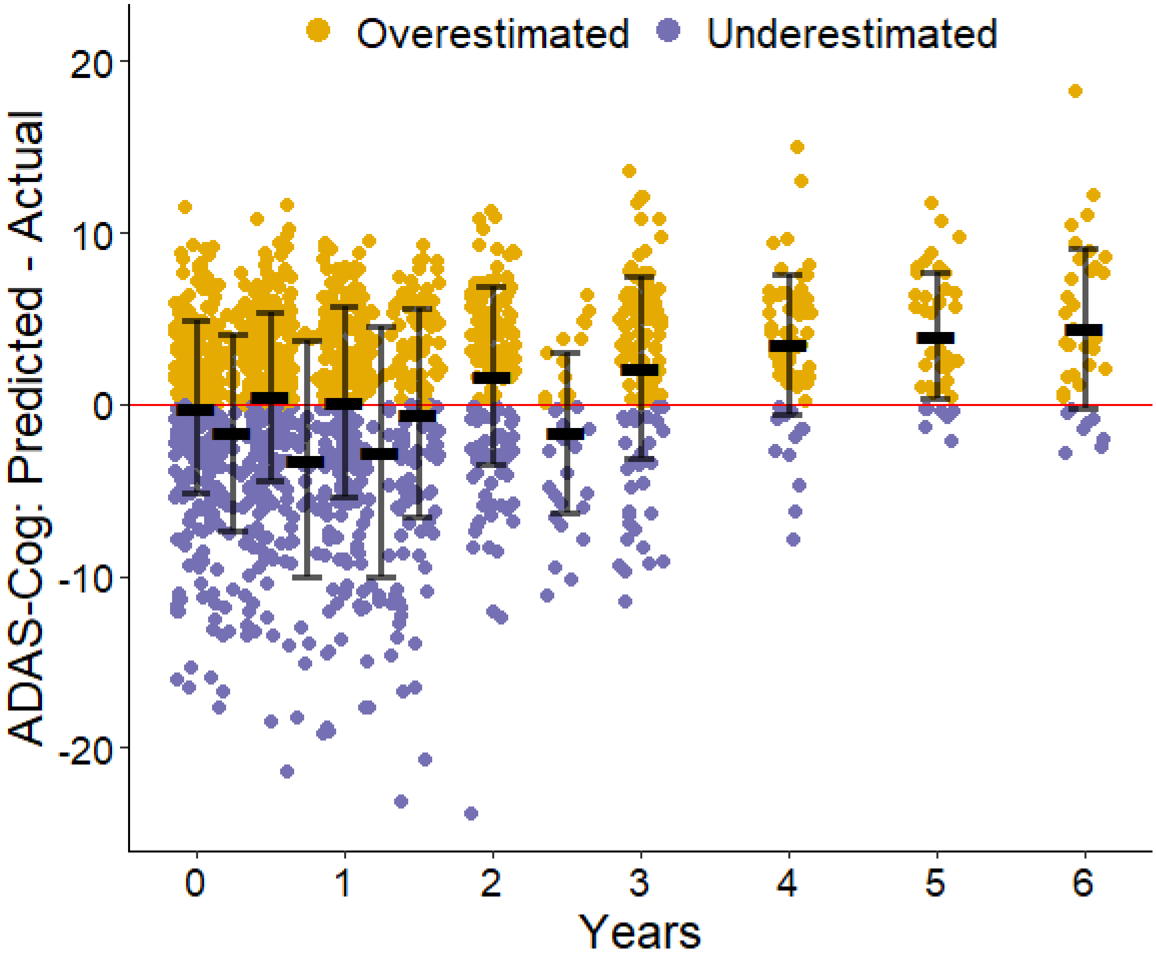
Bagged GLMM Trees Model True and Predicted ADAS-Cog Scores – Whole Subject Trajectories



Note. Error bars centered at mean with standard deviation ranges.

Figure 23

Bagged GLMM Trees Model Prediction Discrepancies for ADAS-Cog Scores – Whole Subject Trajectories



Note. Error bars centered at mean with standard deviation ranges.

Table 19

Bagged GLMM Trees Model Prediction Performance – Final Observation Forecasts of ADAS-Cog Score

Years	RMSE	MAE	Bias	AV Bias
0.25	3.314	3.198	-1.786	3.106
0.5	5.913	3.415	-0.557	1.430
1.0	4.232	3.556	-0.798	2.679
1.25	7.041	5.588	0.319	4.713
1.5	6.515	4.863	-1.138	2.834
2.0	6.460	4.915	-1.074	3.415
3.0	4.068	3.161	-0.092	2.892
4.0	4.854	3.892	0.337	3.011
5.0	2.829	2.180	0.132	1.807
6.0	4.746	3.690	1.597	2.964
All	5.142	3.842	-0.208	2.888

Note. Regression performance metrics presented at individual time points as well as across all time points.

Table 20

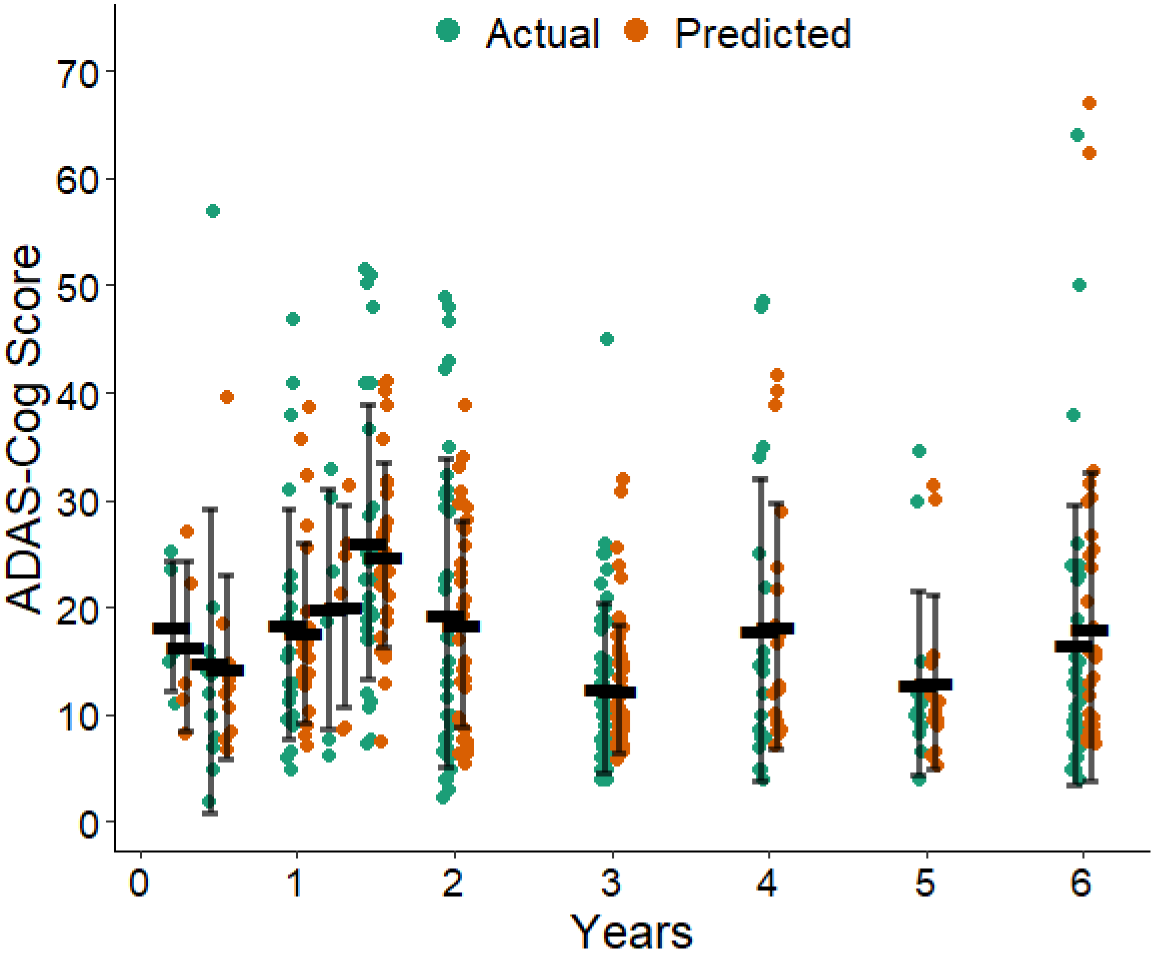
Evaluation of Bagged GLMM Trees Model Relative to CPath Reference – Final Observation Forecasts of ADAS-Cog Score

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
RMSE	5.142	9.720	-4.577	-47.09%	[-5.154, -4.026]
Mean AE	3.842	6.859	-3.017	-43.98%	[-3.464, -2.601]
Bias	-0.208	-1.360	1.152	-84.71%	[0.498, 1.775]
AV bias	2.888	4.799	-1.911	-39.83%	[-2.330, -1.677]

Note. Values for change in metric and percent difference relative to CPath reference model; bootstrap confidence interval corresponds to change in metric.

Figure 24

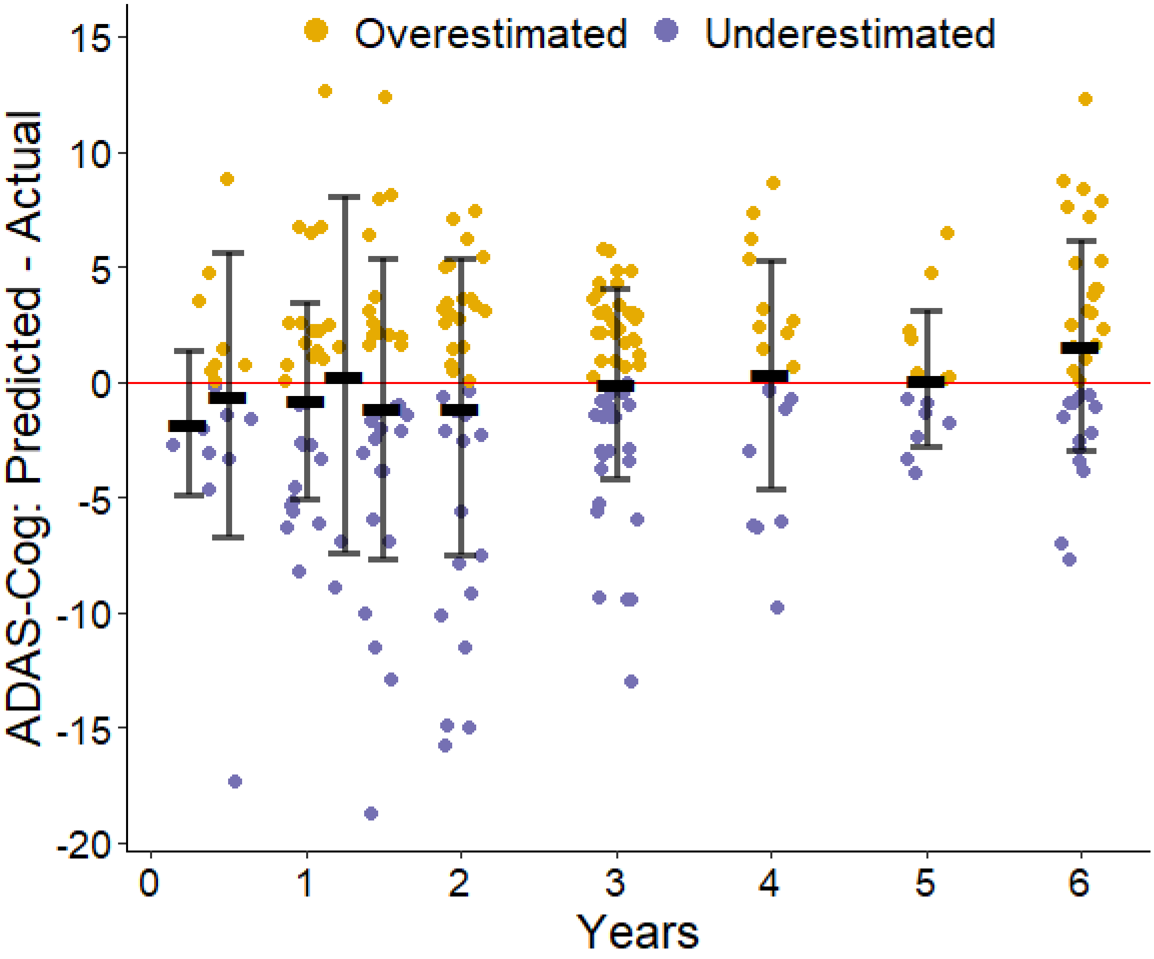
Bagged GLMM Trees Model True and Predicted ADAS-Cog Scores – Final Observation Forecasts



Note. Error bars centered at mean with standard deviation ranges.

Figure 25

Bagged GLMM Trees Model Prediction Discrepancies for ADAS-Cog Scores – Final Observation Forecasts



Note. Error bars centered at mean with standard deviation ranges.

Table 21

Boosted Trees Model Prediction Performance – Whole Subject Trajectories of ADAS-Cog Score

Years	RMSE	MAE	Bias	AV Bias
0.0	4.765	3.642	0.173	2.832
0.25	5.415	4.218	-0.906	3.740
0.5	4.557	3.527	0.604	2.851
0.75	6.204	4.871	-2.741	3.968
1.0	4.936	3.907	0.078	3.273
1.25	6.863	5.508	-2.386	4.346
1.5	5.168	3.801	-0.937	2.936
2.0	5.069	4.160	0.806	3.700
2.5	4.816	3.783	-1.673	2.883
3.0	5.069	4.200	0.787	3.534
4.0	4.015	3.191	1.674	2.550
5.0	3.462	2.726	1.121	2.147
6.0	4.497	3.740	1.370	3.429
All	4.943	3.858	0.107	3.227

Note. Regression performance metrics presented at individual time points as well as across all time points.

Table 22

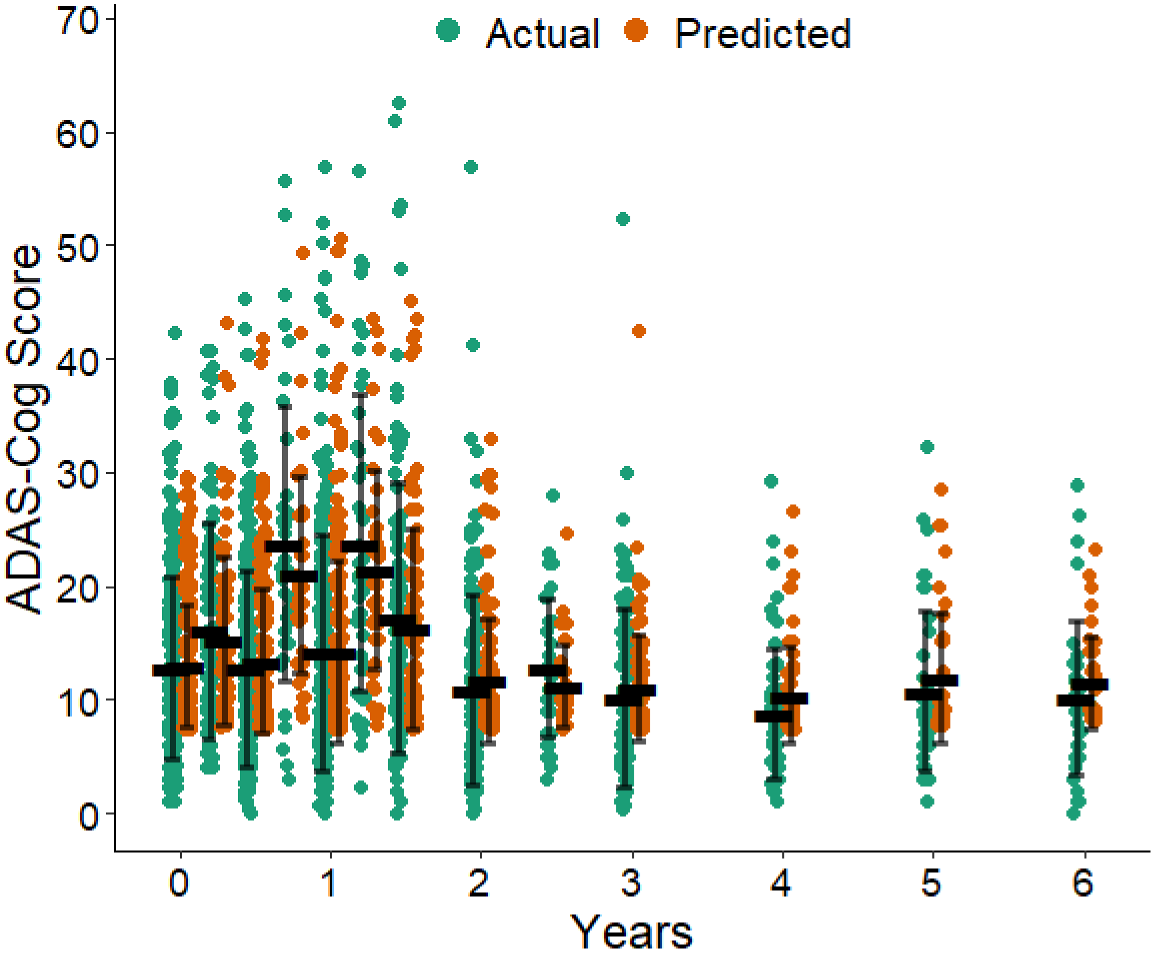
Evaluation of Boosted Trees Model Relative to CPath Reference – Whole Subject Trajectories of ADAS-Cog Score

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
RMSE	4.943	6.819	-1.876	-27.51%	[-2.097, -1.662]
Mean AE	3.858	5.283	-1.426	-26.98%	[-1.590, -1.275]
Bias	0.107	1.602	-1.495	-93.30%	[-1.736, -1.240]
AV bias	3.227	4.421	-1.194	-27.01%	[-1.424, -1.016]

Note. Values for change in metric and percent difference relative to CPath reference model; bootstrap confidence interval corresponds to change in metric.

Figure 26

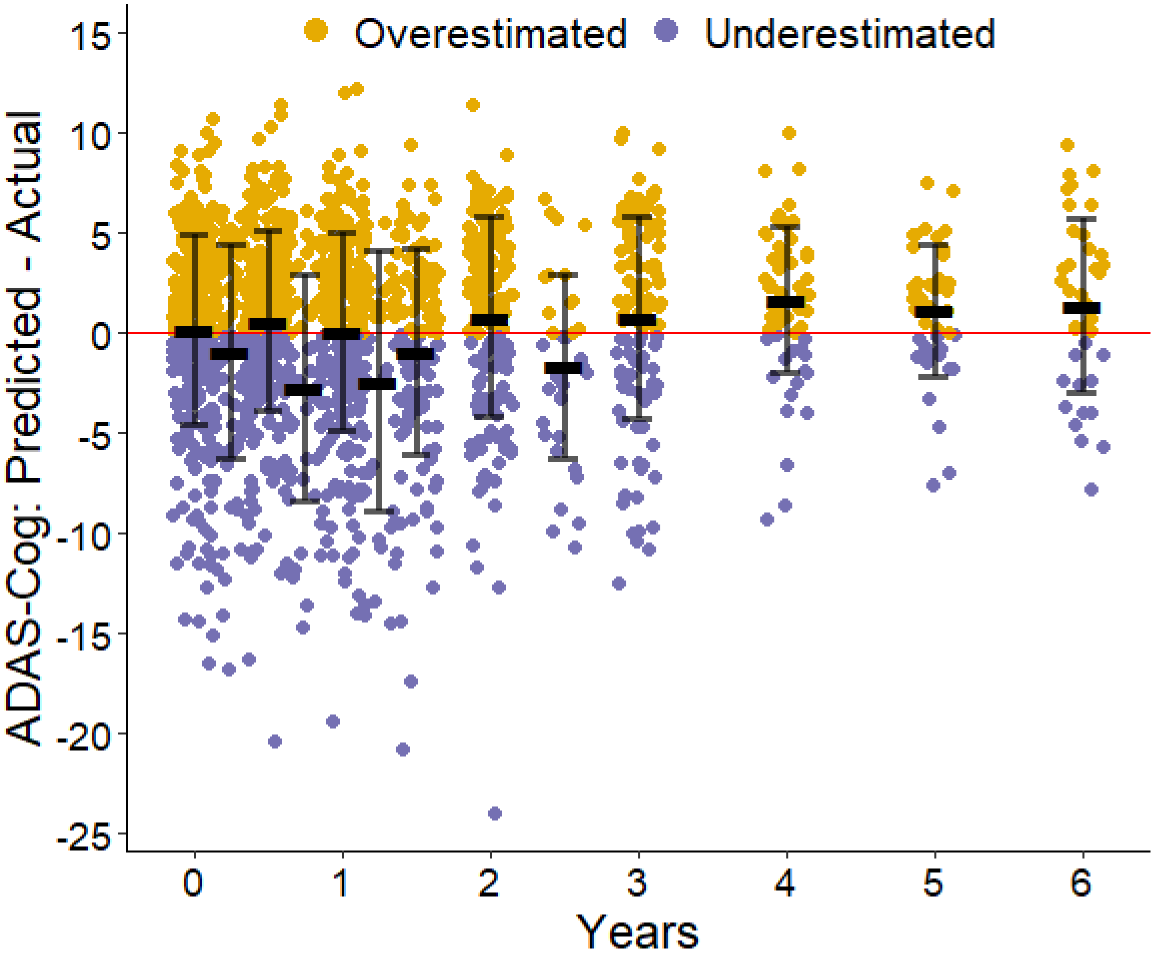
Boosted Trees Model True and Predicted ADAS-Cog Scores – Whole Subject Trajectories



Note. Error bars centered at mean with standard deviation ranges.

Figure 27

Boosted Trees Model Prediction Discrepancies for ADAS-Cog Scores – Whole Subject Trajectories



Note. Error bars centered at mean with standard deviation ranges.

Table 23

Boosted Trees Model Prediction Performance – Final Observation Forecasts of ADAS-Cog Score

Years	RMSE	MAE	Bias	AV Bias
0.25	3.028	2.693	-1.414	2.330
0.5	3.793	2.986	0.072	2.462
1.0	4.110	3.397	-1.297	3.194
1.25	6.506	5.505	-0.149	5.837
1.5	6.779	5.083	-1.998	3.904
2.0	6.969	5.240	-2.160	3.435
3.0	4.317	3.383	-0.541	2.694
4.0	5.280	4.120	-1.803	2.819
5.0	2.905	2.327	-0.990	1.809
6.0	4.726	3.861	-1.651	3.826
All	5.233	3.971	-1.322	3.158

Note. Regression performance metrics presented at individual time points as well as across all time points.

Table 24

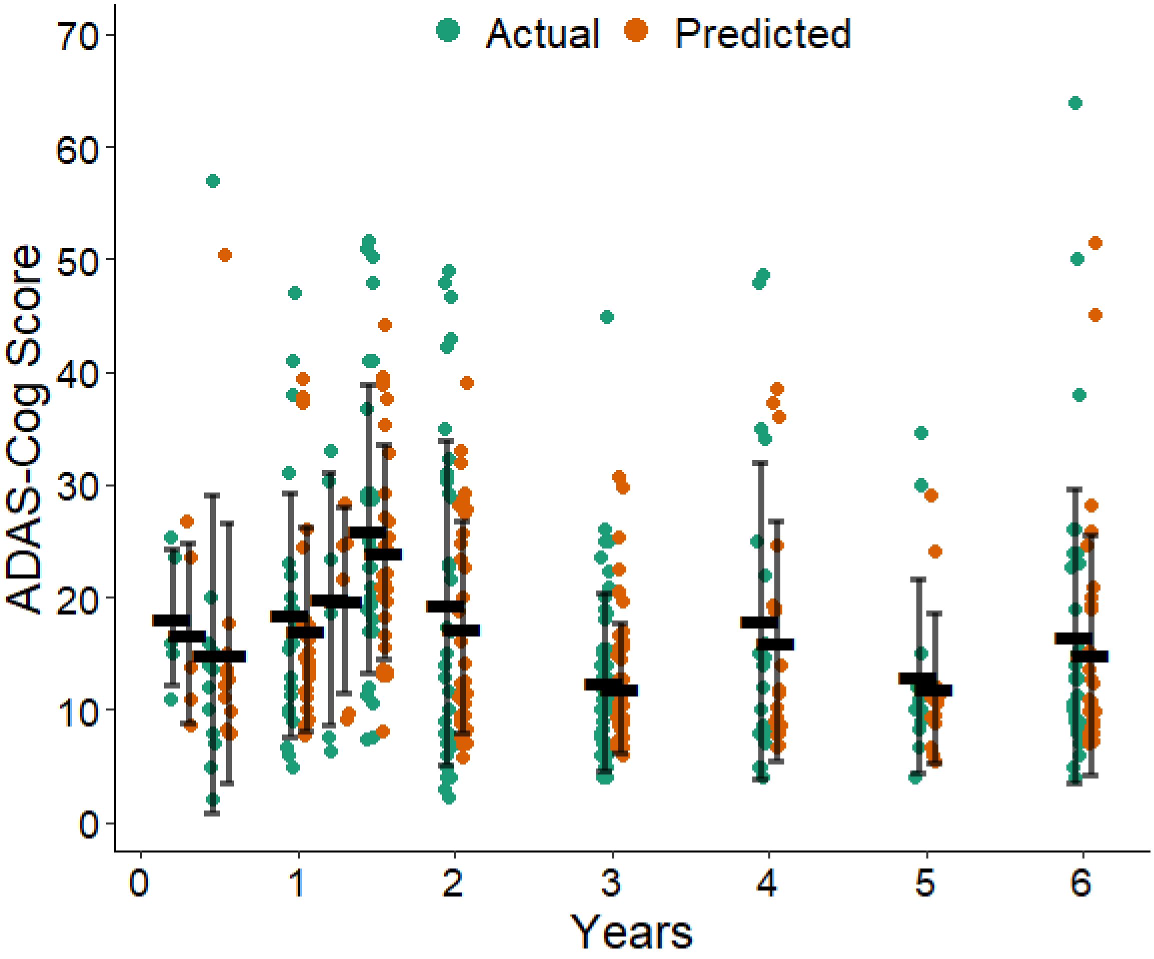
Evaluation of Boosted Trees Model Relative to CPath Reference – Final Observation Forecasts of ADAS-Cog Score

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
RMSE	5.233	9.720	-4.487	-46.16%	[-5.073, -3.937]
Mean AE	3.971	6.859	-2.888	-42.10%	[-3.317, -2.463]
Bias	-1.322	-1.360	0.038	-2.82%	[-0.582, 0.704]
AV bias	3.158	4.799	-1.641	-34.20%	[-2.159, -1.413]

Note. Values for change in metric and percent difference relative to CPath reference model; bootstrap confidence interval corresponds to change in metric.

Figure 28

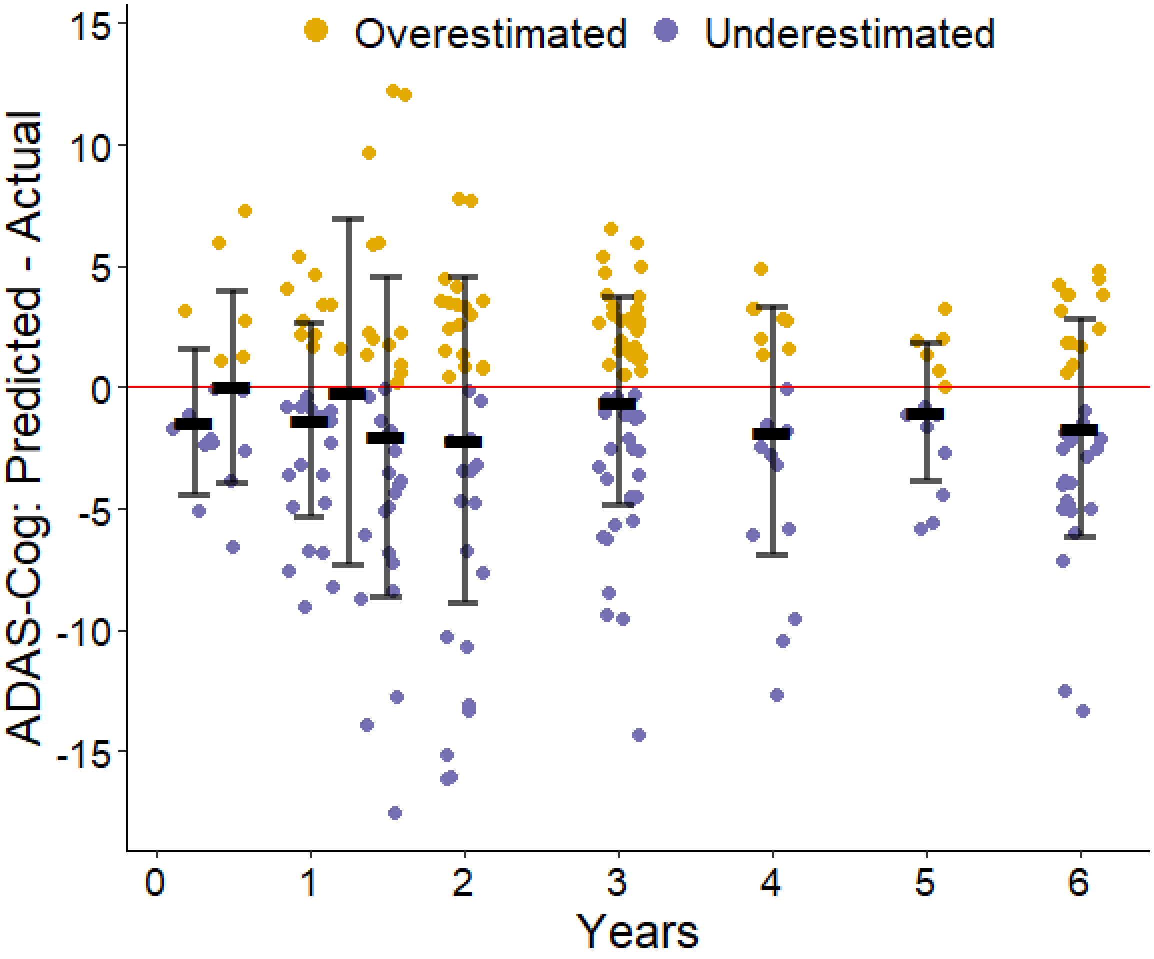
Boosted Trees Model True and Predicted ADAS-Cog Scores – Final Observation Forecasts



Note. Error bars centered at mean with standard deviation ranges.

Figure 29

Boosted Trees Model Prediction Discrepancies for ADAS-Cog Scores – Final Observation Forecasts



Note. Error bars centered at mean with standard deviation ranges.

Table 25

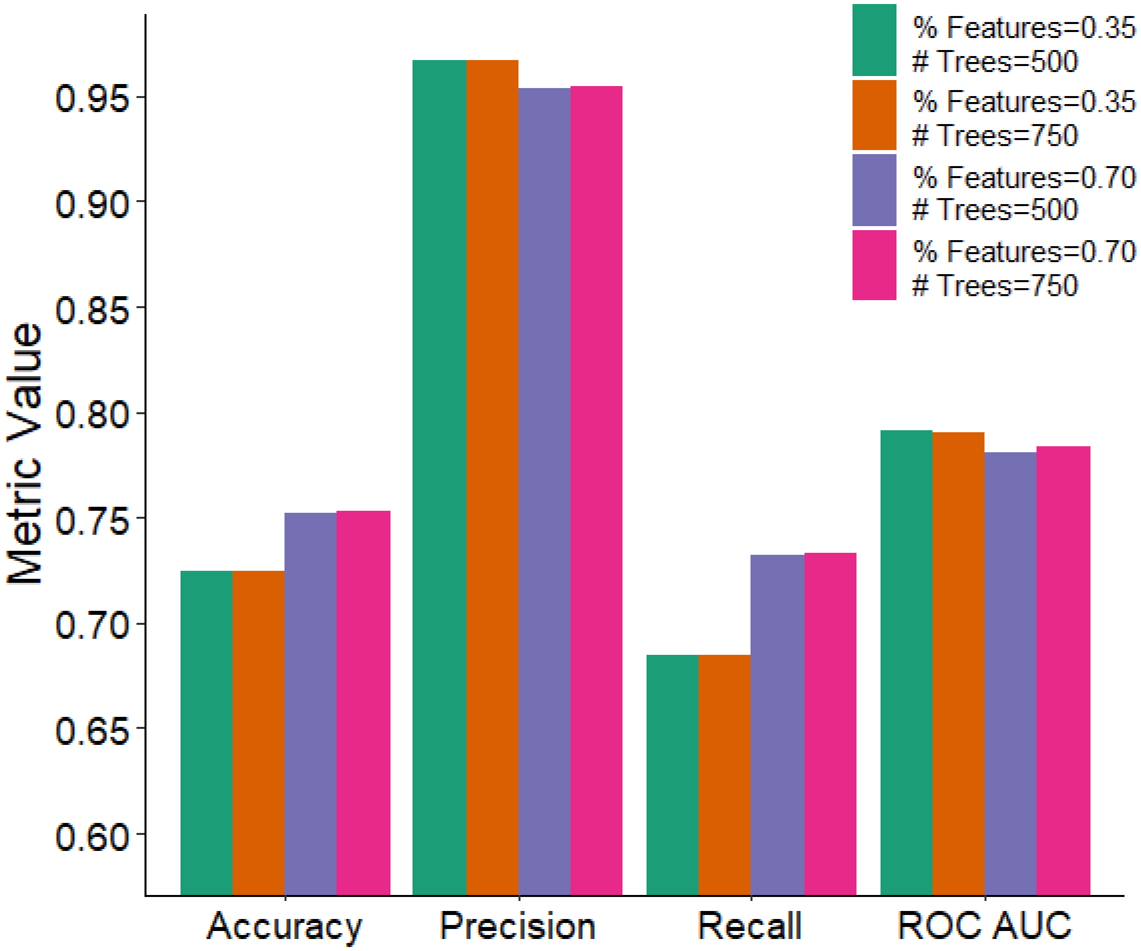
Ensemble Methods Performance Summary – CDR-Based Impairment

Type of prediction	Performance metric	CPath reference	MERF	Single GLMM	Bagged GLMM	Boosted trees
Whole subject trajectories						
	Accuracy	0.711	0.815	0.803	0.818	0.825
	Precision	0.958	0.960	0.967	0.960	0.965
	Recall	0.661	0.798	0.776	0.801	0.806
	ROC AUC	0.841	0.903	0.892	0.907	0.916
Final observation forecasts						
	Accuracy	0.730	0.967	0.967	0.963	0.935
	Precision	0.977	0.983	0.989	0.989	0.982
	Recall	0.694	0.978	0.972	0.966	0.939
	ROC AUC	0.825	0.981	0.981	0.987	0.972

Note. Classification performance metrics summarized across all time points with logistic model shown for reference.

Figure 30

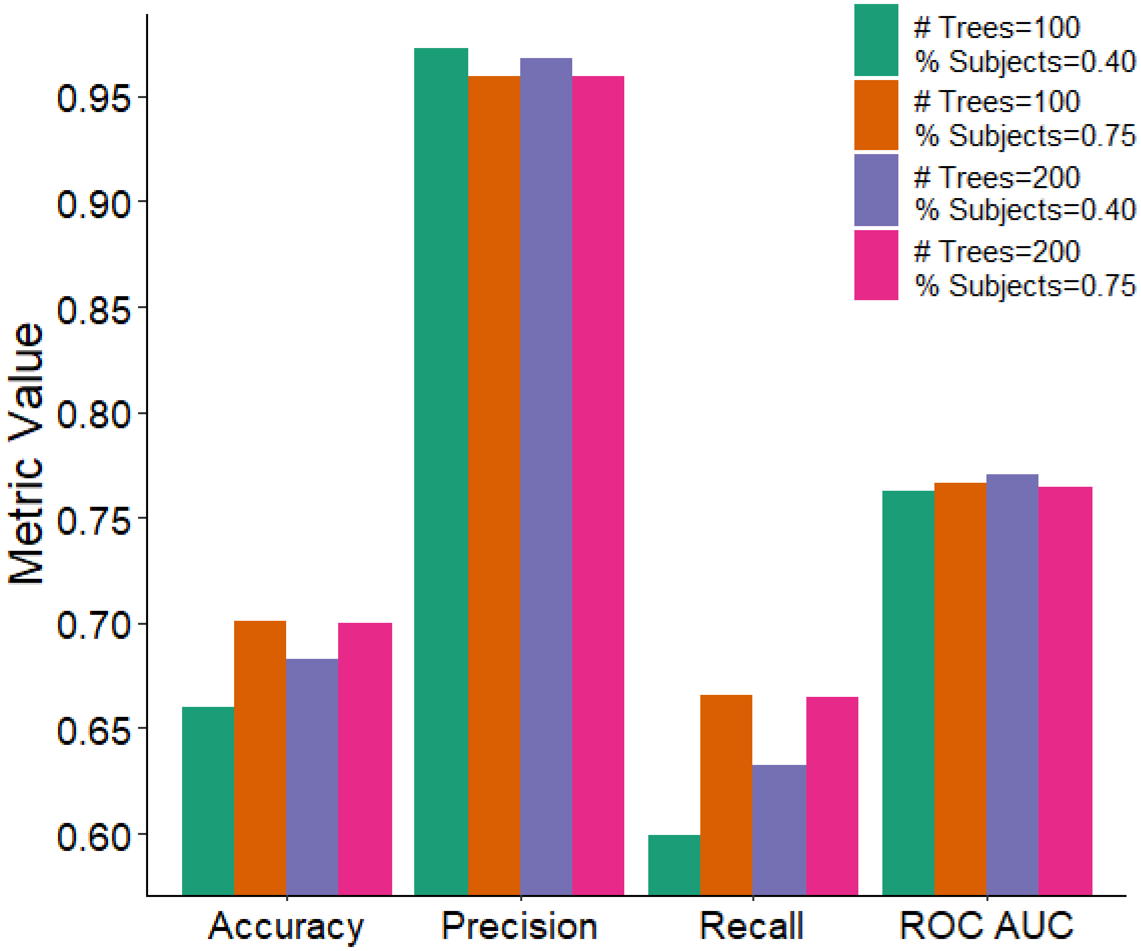
MERF Model Hyperparameter Tunings – CDR-Based Impairment



Note. Hyperparameter tunings provided from cross-fold validations on whole subject trajectories.

Figure 31

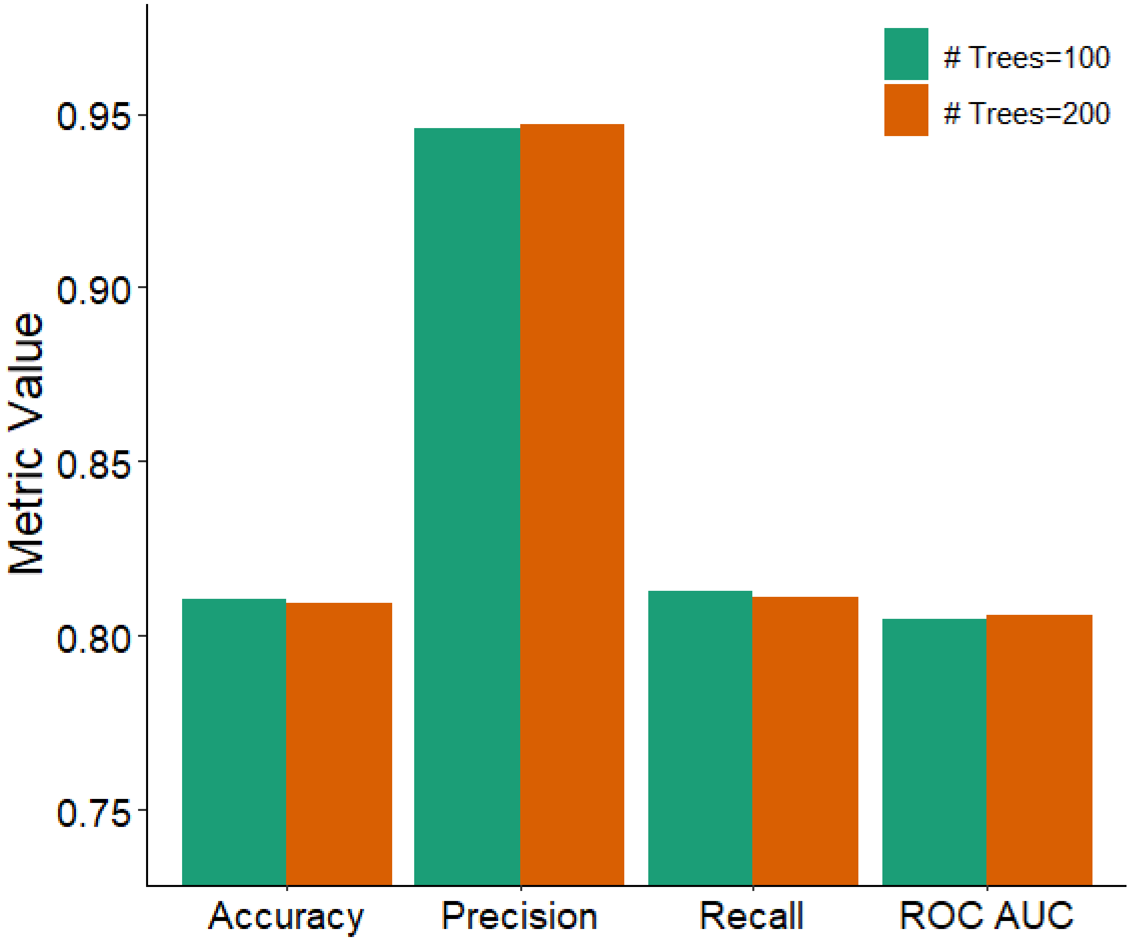
Bagged GLMM Trees Model Hyperparameter Tunings – CDR-Based Impairment



Note. Hyperparameter tunings provided from cross-fold validations on whole subject trajectories.

Figure 32

Boosted Trees Model Hyperparameter Tunings – CDR-Based Impairment



Note. Hyperparameter tunings provided from cross-fold validations on whole subject trajectories.

Table 26

Top Hyperparameters for Ensemble Methods – CDR-Based Impairment

Ensemble method	Performance metric	Best value	Hyperparameter set
MERF model			
	Accuracy	0.754	% Features=0.70 ; # Trees=750
	Precision	0.967	% Features=0.35 ; # Trees=500
	Recall	0.733	% Features=0.70 ; # Trees=750
	ROC AUC	0.791	% Features=0.35 ; # Trees=500
Bagged GLMM trees model			
	Accuracy	0.701	# Trees=100 ; % Subjects=0.75
	Precision	0.973	# Trees=100 ; % Subjects=0.40
	Recall	0.665	# Trees=100 ; % Subjects=0.75
	ROC AUC	0.770	# Trees=200 ; % Subjects=0.40
Boosted trees model			
	Accuracy	0.811	# Trees=100
	Precision	0.947	# Trees=200
	Recall	0.813	# Trees=100
	ROC AUC	0.806	# Trees=200

Note. For each ensemble method, the best performing hyperparameter set for each classification metric is displayed along with the corresponding value.

Table 27

MERF Model Prediction Performance – Whole Subject Trajectories of CDR-Based Impairment

Years	Accuracy	Precision	Recall	AUC	NRI
0.0	0.859	0.977	0.849	0.932	0.108
0.5	0.847	0.972	0.838	0.911	0.105
1.0	0.829	0.957	0.830	0.899	0.078
1.5	0.839	0.959	0.869	0.802	-0.288
2.0	0.755	0.934	0.703	0.867	0.139
2.5	0.655	1.000	0.655	1.000	-
3.0	0.663	0.907	0.609	0.809	0.080
4.0	0.824	0.923	0.600	0.847	0.282
5.0	0.788	0.917	0.647	0.934	0.357
6.0	0.812	0.900	0.643	0.861	0.468
All	0.815	0.960	0.798	0.903	0.124

Note. Classification performance metrics presented at individual time points as well as across all time points. NRI based on reclassification from logistic reference model; NRI incalculable under certain ROC AUC conditions.

Table 28

Evaluation of MERF Model Relative to Logistic Reference – Whole Subject Trajectories of CDR-Based Impairment

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
Accuracy	0.815	0.711	0.105	14.72%	[0.082, 0.125]
Precision	0.960	0.958	0.002	0.23%	[-0.011, 0.015]
Recall	0.798	0.661	0.138	20.82%	[0.111, 0.161]
ROC AUC	0.903	0.841	0.061	7.30%	[0.043, 0.078]

Note. Values for change in metric and percent difference relative to logistic reference model; bootstrap confidence interval corresponds to change in metric.

Figure 33

MERF Model True and Predicted CDR-Based Impairment Counts – Whole Subject Trajectories

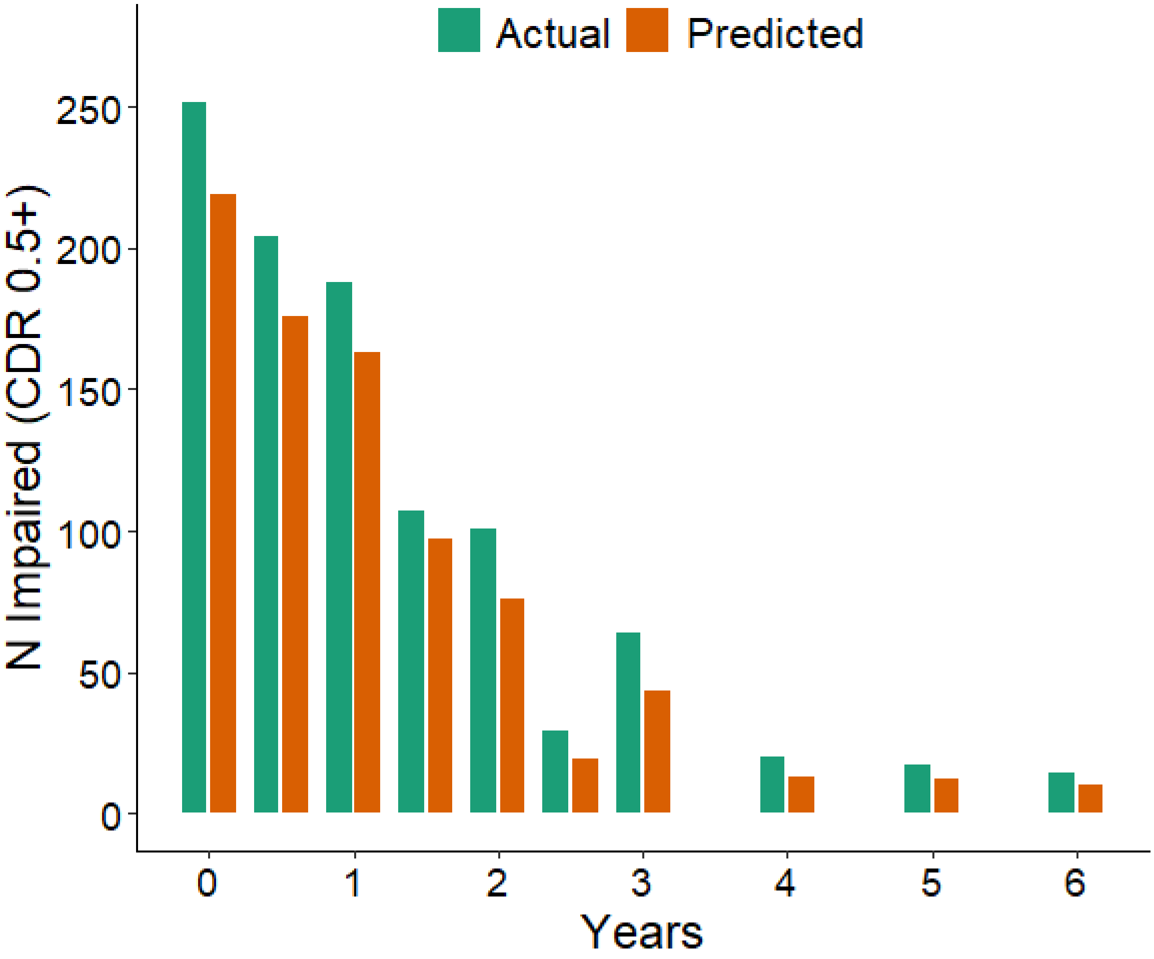
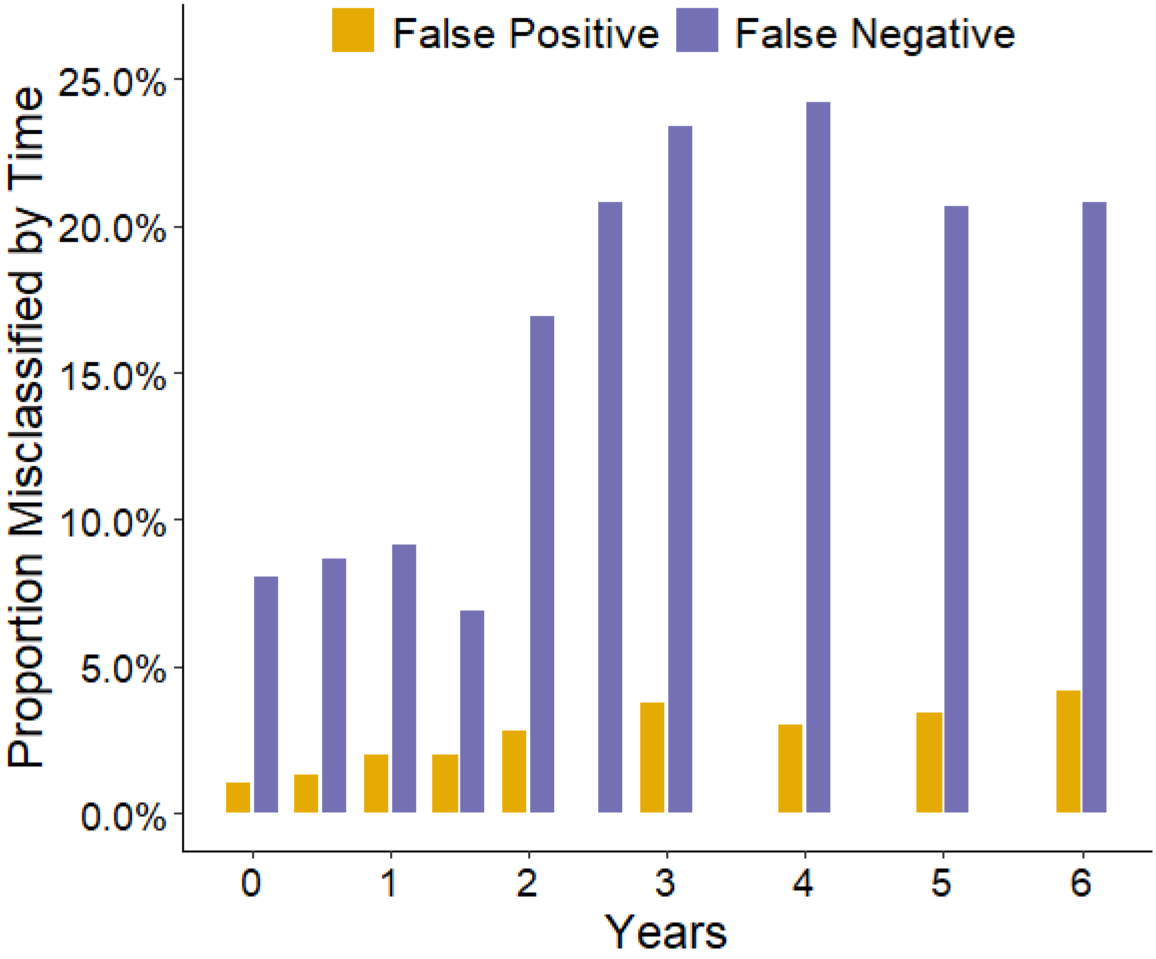


Figure 34

MERF Model Misclassification Rates for CDR-Based Impairment – Whole Subject Trajectories



Note. False positive and false negative rates marginalized by time to be relative to total counts at that timepoint.

Table 29

MERF Model Prediction Performance – Final Observation Forecasts of CDR-Based Impairment

Years	Accuracy	Precision	Recall	AUC	NRI
0.5	0.929	1.000	0.923	1.000	1.308
1.0	1.000	1.000	1.000	1.000	-
1.5	1.000	1.000	1.000	1.000	-
2.0	1.000	1.000	1.000	1.000	-
3.0	0.957	0.975	0.975	0.971	0.425
4.0	0.933	0.909	1.000	0.940	0.300
5.0	0.917	1.000	0.857	0.971	0.143
6.0	0.935	0.947	0.947	0.987	0.443
All	0.967	0.983	0.978	0.981	0.285

Note. Classification performance metrics presented at individual time points as well as across all time points. NRI based on reclassification from logistic reference model; NRI incalculable under certain ROC AUC conditions.

Table 30

Evaluation of MERF Model Relative to Logistic Reference – Final Observation Forecasts of CDR-Based Impairment

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
Accuracy	0.967	0.730	0.237	32.46%	[0.209, 0.256]
Precision	0.983	0.977	0.007	0.67%	[-0.016, 0.023]
Recall	0.978	0.694	0.283	40.78%	[0.260, 0.300]
ROC AUC	0.981	0.825	0.157	19.01%	[0.132, 0.172]

Note. Values for change in metric and percent difference relative to logistic reference model; bootstrap confidence interval corresponds to change in metric.

Figure 35

MERF Model True and Predicted CDR-Based Impairment Counts – Final Observation Forecasts

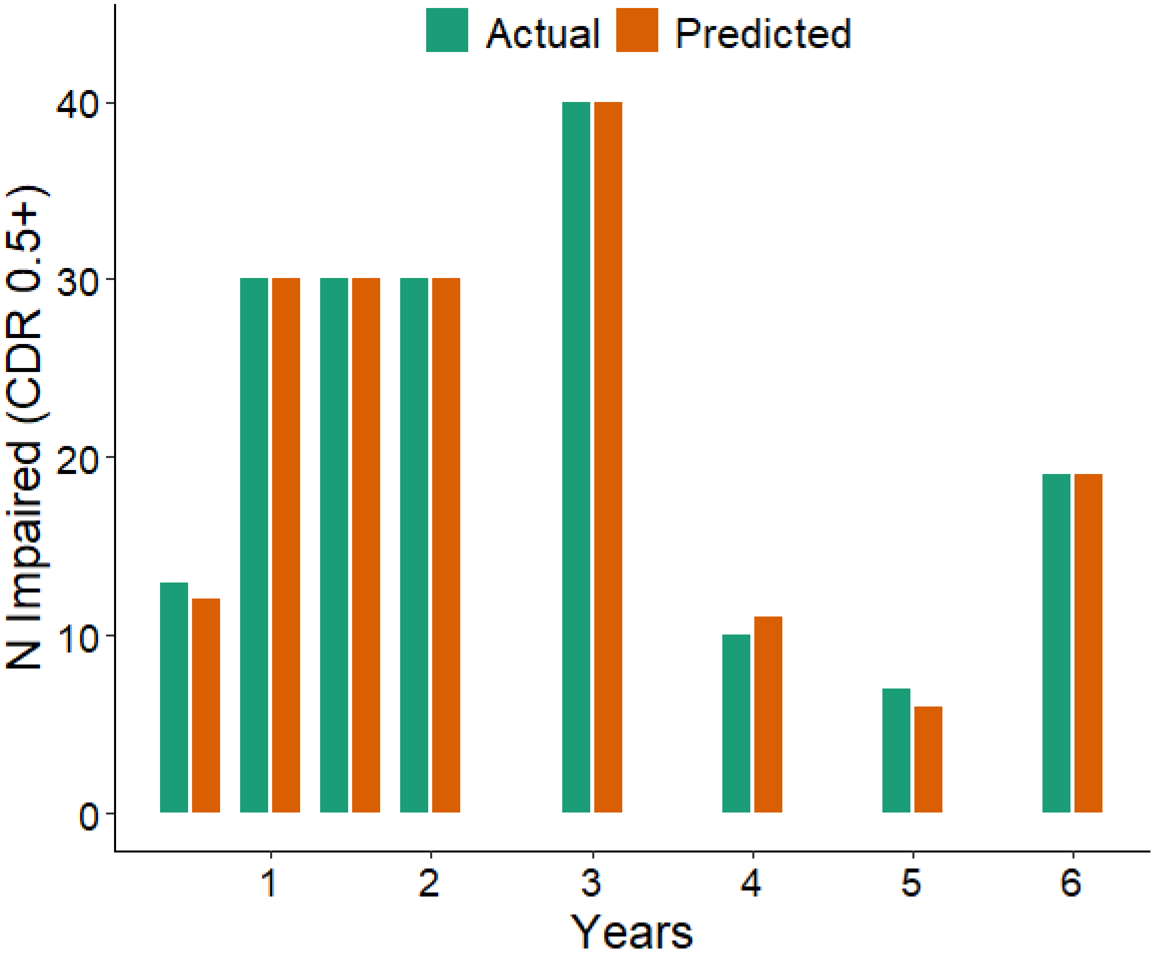
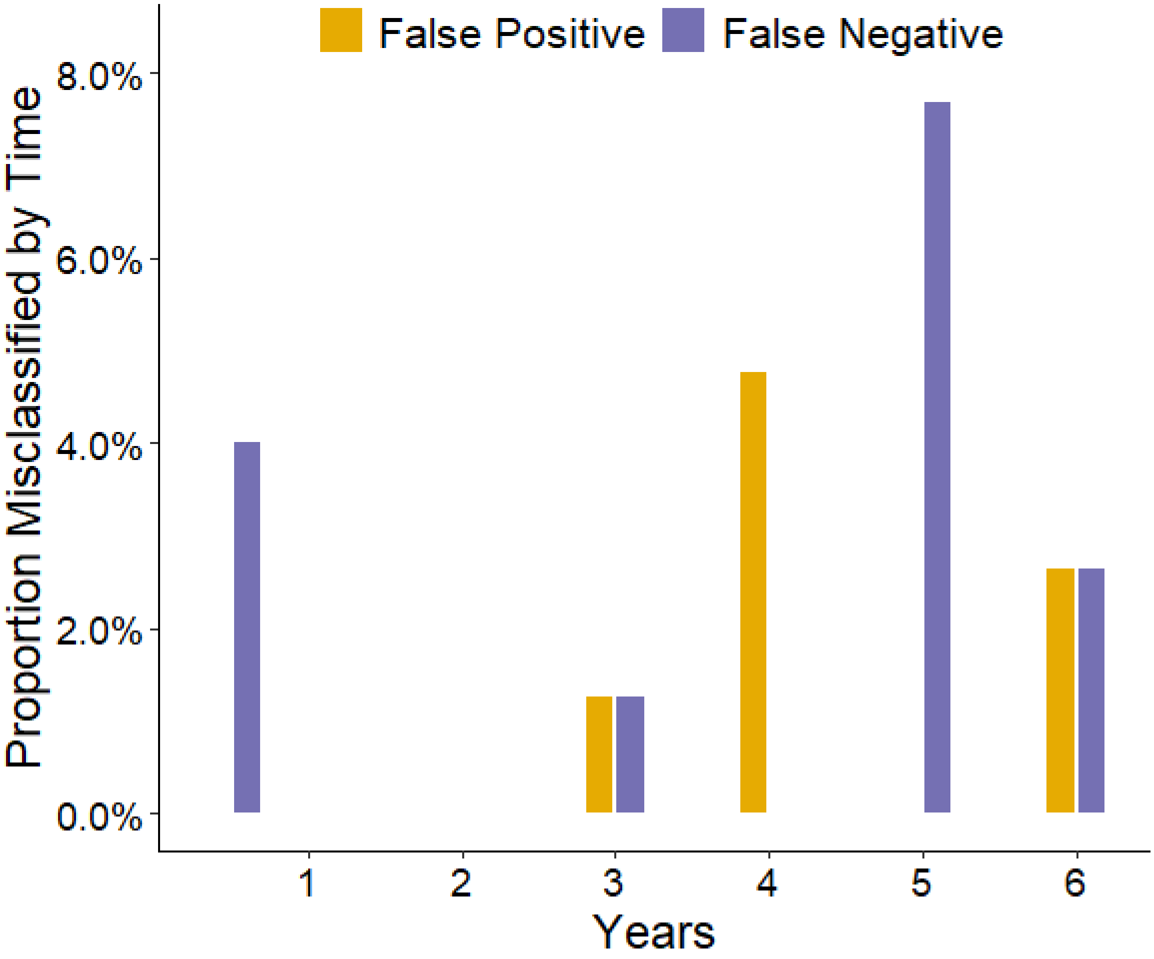


Figure 36

MERF Model Misclassification Rates for CDR-Based Impairment – Final Observation Forecasts



Note. False positive and false negative rates marginalized by time to be relative to total counts at that timepoint.

Table 31

Single GLMM Tree Model Prediction Performance – Whole Subject Trajectories of CDR-Based Impairment

Years	Accuracy	Precision	Recall	AUC	NRI
0.0	0.862	0.977	0.853	0.923	0.112
0.5	0.823	0.960	0.819	0.891	0.041
1.0	0.825	0.981	0.803	0.901	0.151
1.5	0.804	0.978	0.813	0.815	0.056
2.0	0.755	0.958	0.683	0.871	0.166
2.5	0.414	1.000	0.414	1.000	-
3.0	0.698	0.952	0.625	0.791	0.186
4.0	0.824	0.923	0.600	0.818	0.282
5.0	0.818	0.923	0.706	0.908	0.415
6.0	0.750	0.800	0.571	0.831	0.341
All	0.803	0.967	0.776	0.892	0.128

Note. Classification performance metrics presented at individual time points as well as across all time points. NRI based on reclassification from logistic reference model; NRI incalculable under certain ROC AUC conditions.

Table 32

Evaluation of Single GLMM Tree Model Relative to Logistic Reference – Whole Subject Trajectories of CDR-Based Impairment

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
Accuracy	0.803	0.711	0.093	13.06%	[0.071, 0.114]
Precision	0.967	0.958	0.010	1.00%	[-0.003, 0.021]
Recall	0.776	0.661	0.115	17.47%	[0.089, 0.141]
ROC AUC	0.892	0.841	0.051	6.09%	[0.031, 0.068]

Note. Values for change in metric and percent difference relative to logistic reference model; bootstrap confidence interval corresponds to change in metric.

Figure 37

Single GLMM Tree Model True and Predicted CDR-Based Impairment Counts – Whole Subject Trajectories

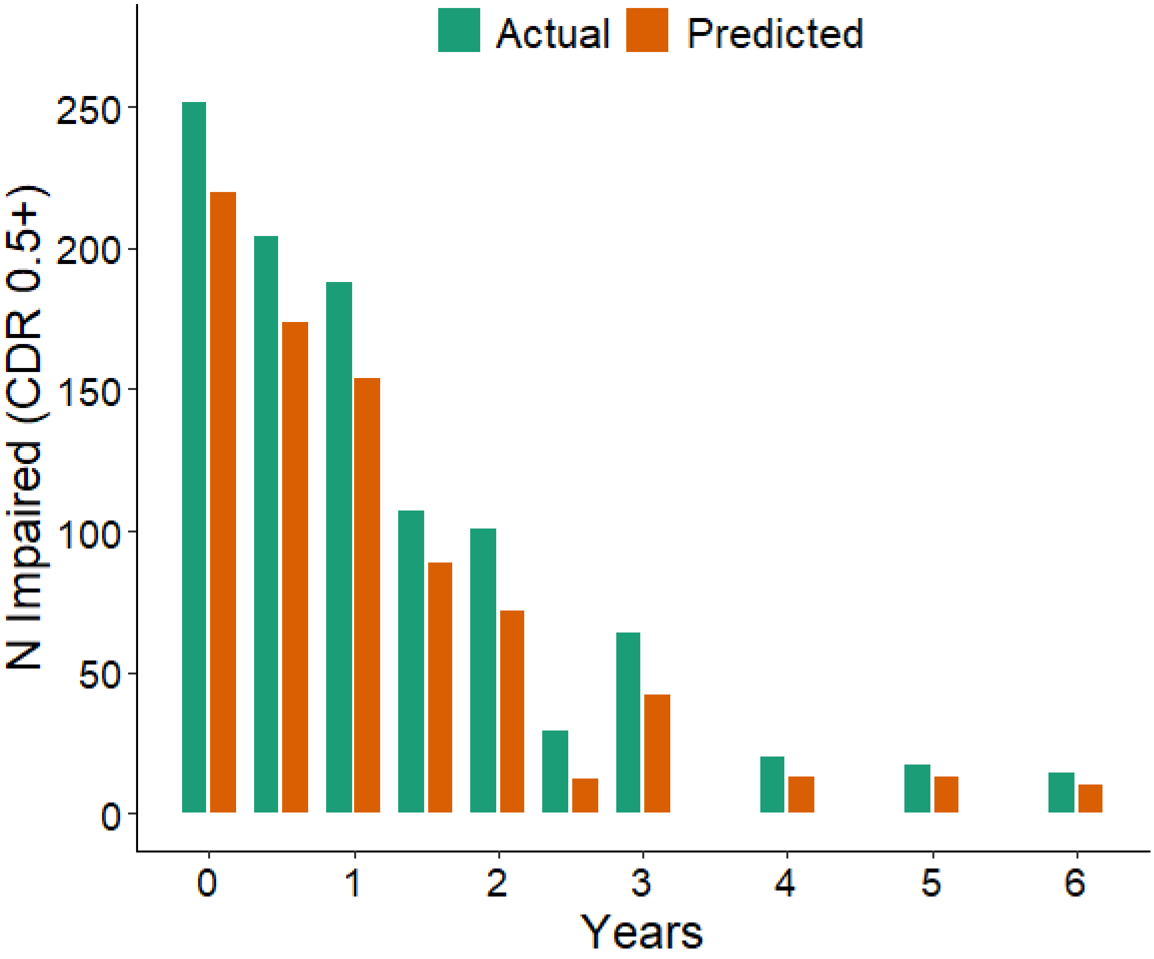
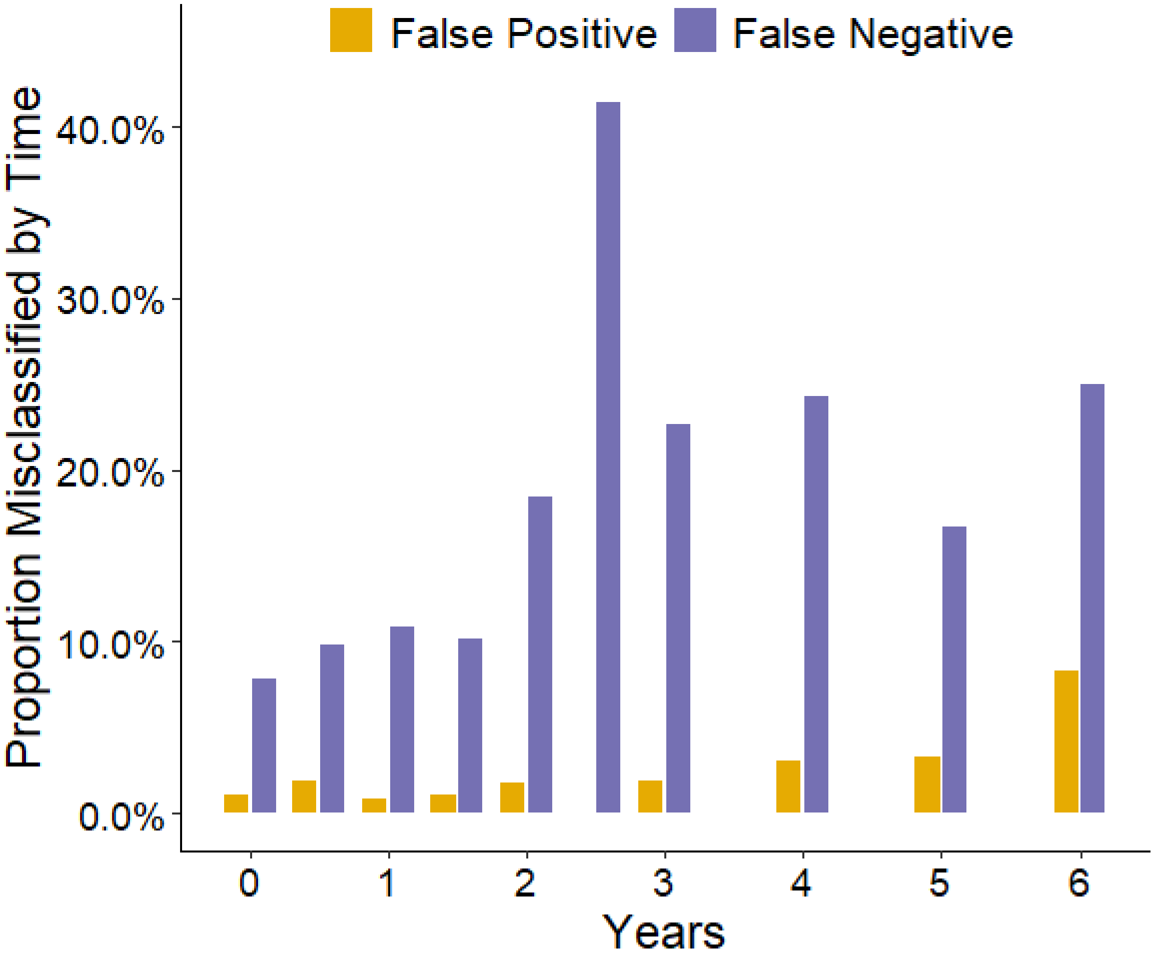


Figure 38

Single GLMM Tree Model Misclassification Rates for CDR-Based Impairment – Whole Subject Trajectories



Note. False positive and false negative rates marginalized by time to be relative to total counts at that timepoint.

Table 33

Single GLMM Tree Model Prediction Performance – Final Observation Forecasts of CDR-Based Impairment

Years	Accuracy	Precision	Recall	AUC	NRI
0.5	0.929	1.000	0.923	1.000	1.308
1.0	1.000	1.000	1.000	1.000	-
1.5	1.000	1.000	1.000	1.000	-
2.0	0.972	1.000	0.967	1.000	-
3.0	0.957	0.975	0.975	0.917	0.425
4.0	0.933	0.909	1.000	0.980	0.300
5.0	0.917	1.000	0.857	1.000	0.143
6.0	0.968	1.000	0.947	0.982	0.526
All	0.967	0.989	0.972	0.981	0.308

Note. Classification performance metrics presented at individual time points as well as across all time points. NRI based on reclassification from logistic reference model; NRI incalculable under certain ROC AUC conditions.

Table 34

Evaluation of Single GLMM Tree Model Relative to Logistic Reference – Final Observation Forecasts of CDR-Based Impairment

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
Accuracy	0.967	0.730	0.237	32.46%	[0.214, 0.260]
Precision	0.989	0.977	0.012	1.24%	[-0.006, 0.023]
Recall	0.972	0.694	0.278	39.98%	[0.254, 0.300]
ROC AUC	0.981	0.825	0.157	19.01%	[0.133, 0.173]

Note. Values for change in metric and percent difference relative to logistic reference model; bootstrap confidence interval corresponds to change in metric.

Figure 39

Single GLMM Tree Model True and Predicted CDR-Based Impairment Counts – Final Observation Forecasts

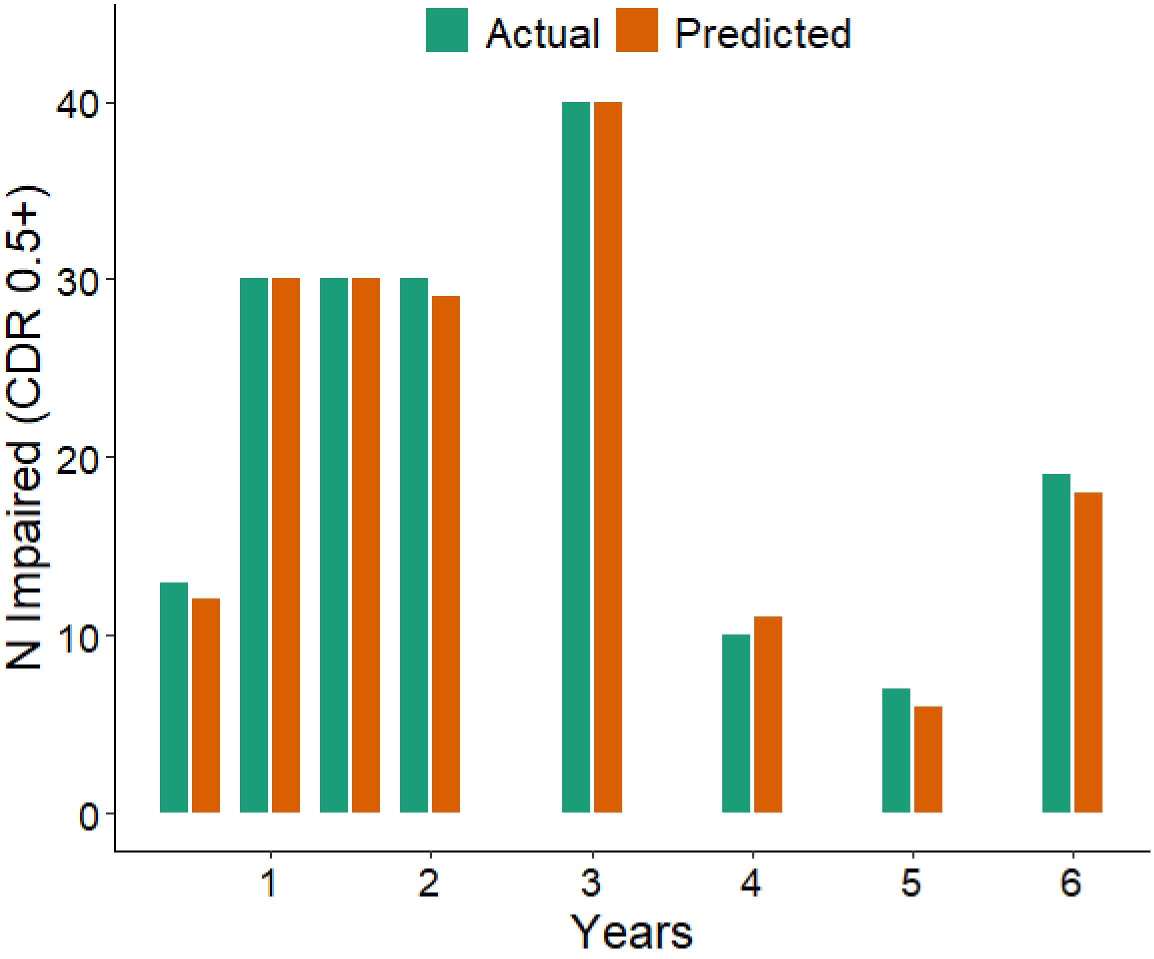
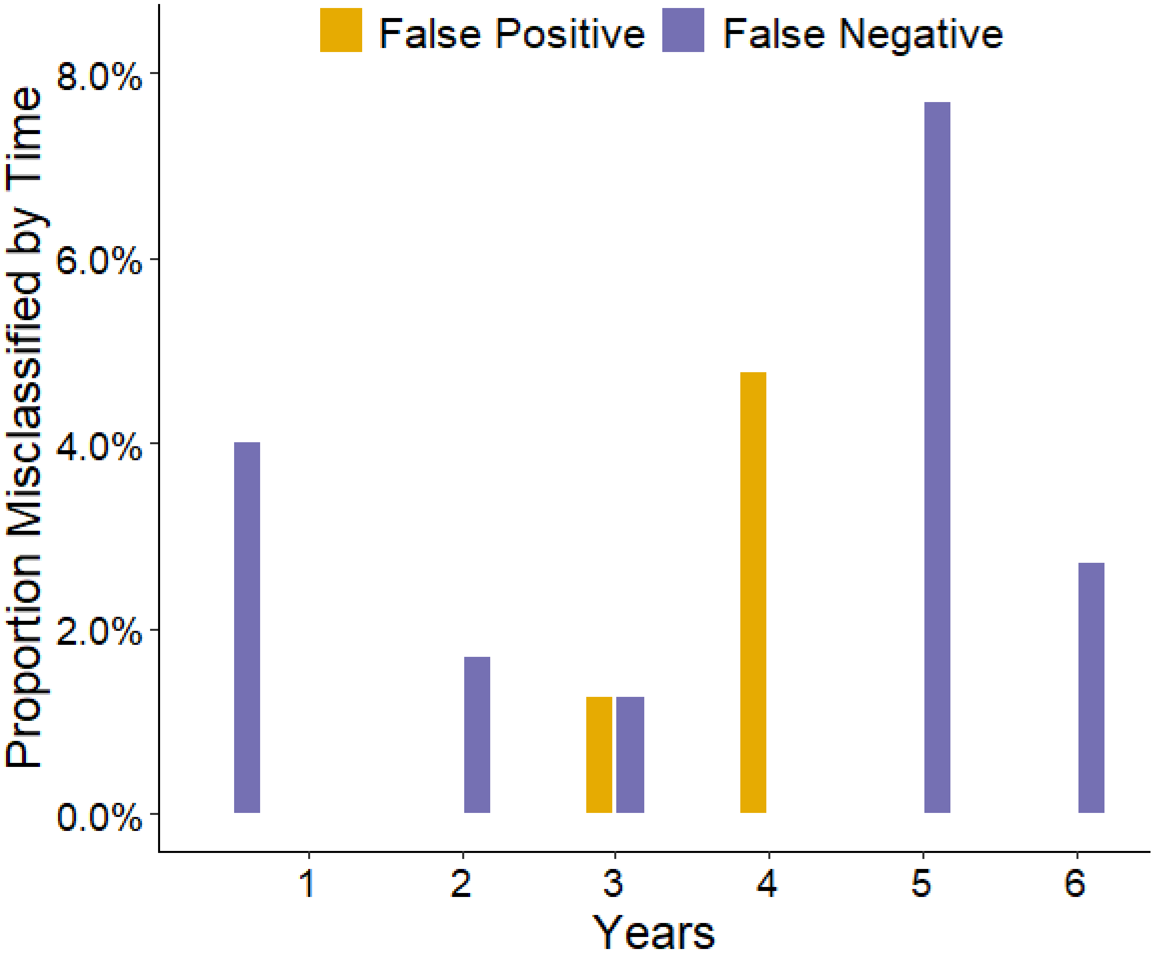


Figure 40

Single GLMM Tree Model Misclassification Rates for CDR-Based Impairment – Final Observation Forecasts



Note. False positive and false negative rates marginalized by time to be relative to total counts at that timepoint.

Table 35

Bagged GLMM Trees Model Prediction Performance – Whole Subject Trajectories of CDR-Based Impairment

Years	Accuracy	Precision	Recall	AUC	NRI
0.0	0.868	0.973	0.865	0.929	0.105
0.5	0.835	0.960	0.833	0.913	0.056
1.0	0.829	0.963	0.824	0.910	0.097
1.5	0.821	0.958	0.850	0.755	-0.307
2.0	0.797	0.962	0.743	0.887	0.226
2.5	0.552	1.000	0.552	1.000	-
3.0	0.674	0.909	0.625	0.824	0.095
4.0	0.843	0.929	0.650	0.852	0.332
5.0	0.788	0.917	0.647	0.923	0.357
6.0	0.812	0.900	0.643	0.917	0.468
All	0.818	0.960	0.801	0.907	0.127

Note. Classification performance metrics presented at individual time points as well as across all time points. NRI based on reclassification from logistic reference model; NRI incalculable under certain ROC AUC conditions.

Table 36

Evaluation of Bagged GLMM Trees Model Relative to Logistic Reference – Whole Subject Trajectories of CDR-Based Impairment

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
Accuracy	0.818	0.711	0.107	15.05%	[0.086, 0.126]
Precision	0.960	0.958	0.002	0.25%	[-0.011, 0.015]
Recall	0.801	0.661	0.141	21.27%	[0.115, 0.163]
ROC AUC	0.907	0.841	0.066	7.80%	[0.047, 0.082]

Note. Values for change in metric and percent difference relative to logistic reference model; bootstrap confidence interval corresponds to change in metric.

Figure 41

Bagged GLMM Trees Model True and Predicted CDR-Based Impairment Counts – Whole Subject Trajectories

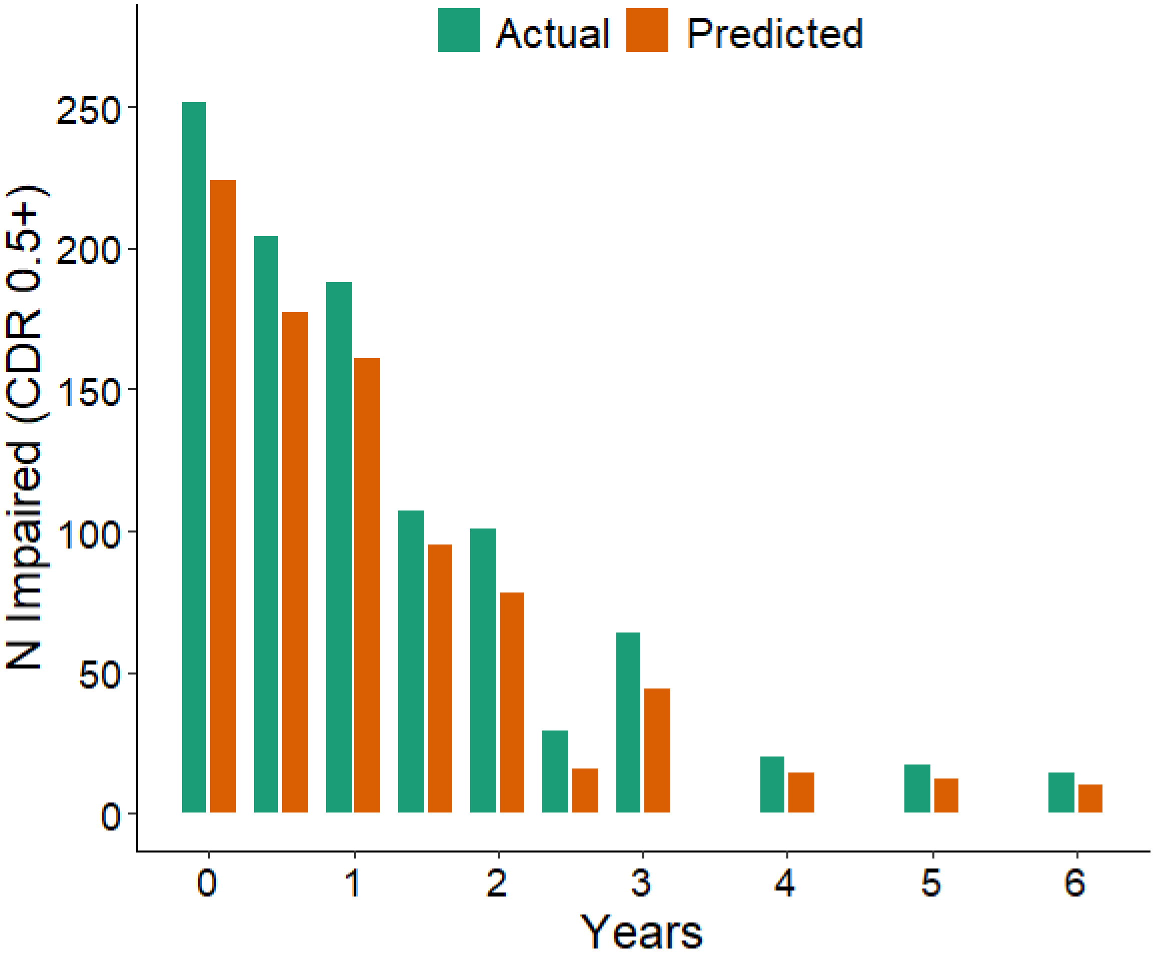
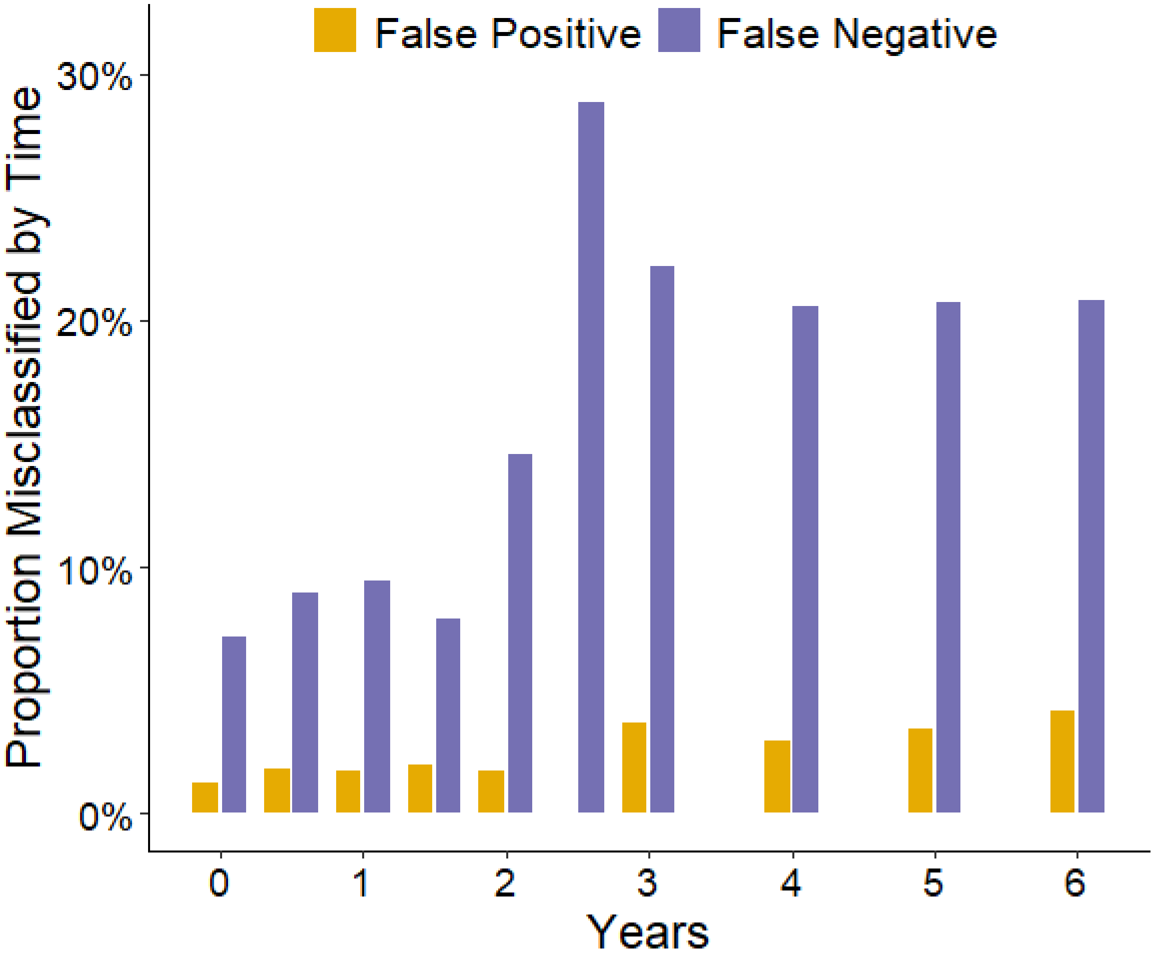


Figure 42

Bagged GLMM Trees Model Misclassification Rates for CDR-Based Impairment – Whole Subject Trajectories



Note. False positive and false negative rates marginalized by time to be relative to total counts at that timepoint.

Table 37

Bagged GLMM Trees Model Prediction Performance – Final Observation Forecasts of CDR-Based Impairment

Years	Accuracy	Precision	Recall	AUC	NRI
0.5	0.929	1.000	0.923	1.000	1.308
1.0	1.000	1.000	1.000	1.000	-
1.5	1.000	1.000	1.000	1.000	-
2.0	0.972	1.000	0.967	1.000	0.333
3.0	0.957	0.975	0.975	0.950	0.425
4.0	0.867	0.900	0.900	0.980	0.200
5.0	0.917	1.000	0.857	0.971	0.143
6.0	0.968	1.000	0.947	0.982	0.526
All	0.963	0.989	0.966	0.987	0.302

Note. Classification performance metrics presented at individual time points as well as across all time points. NRI based on reclassification from logistic reference model; NRI incalculable under certain ROC AUC conditions.

Table 38

Evaluation of Bagged GLMM Trees Model Relative to Logistic Reference – Final Observation Forecasts of CDR-Based Impairment

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
Accuracy	0.963	0.730	0.232	31.82%	[0.204, 0.256]
Precision	0.989	0.977	0.012	1.23%	[-0.005, 0.023]
Recall	0.966	0.694	0.272	39.17%	[0.243, 0.295]
ROC AUC	0.987	0.825	0.162	19.66%	[0.148, 0.173]

Note. Values for change in metric and percent difference relative to logistic reference model; bootstrap confidence interval corresponds to change in metric.

Figure 43

Bagged GLMM Trees Model True and Predicted CDR-Based Impairment Counts – Final Observation Forecasts

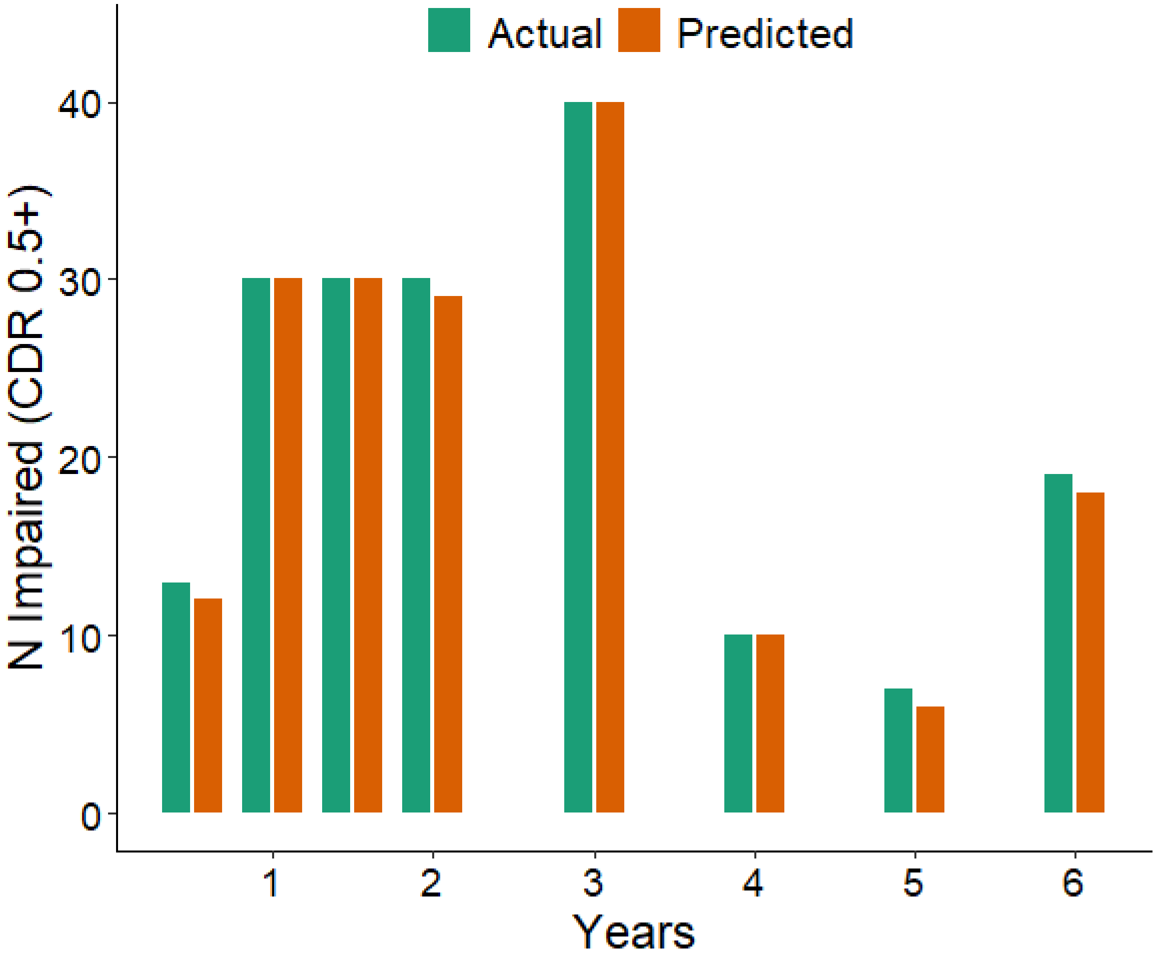
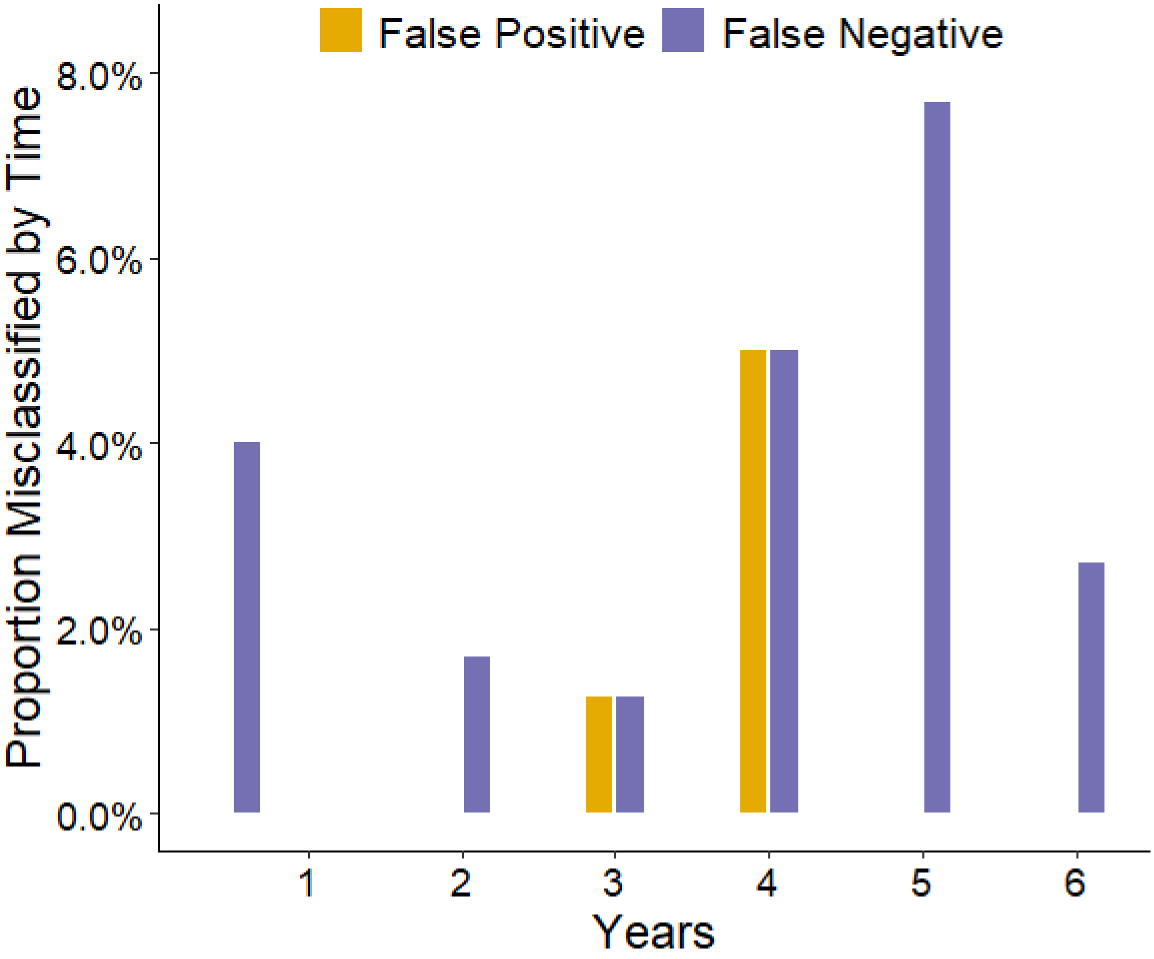


Figure 44

Bagged GLMM Trees Model Misclassification Rates for CDR-Based Impairment – Final Observation Forecasts



Note. False positive and false negative rates marginalized by time to be relative to total counts at that timepoint.

Table 39

Boosted Trees Model Prediction Performance – Whole Subject Trajectories of CDR-Based Impairment

Years	Accuracy	Precision	Recall	AUC	NRI
0.0	0.859	0.982	0.845	0.936	0.124
0.5	0.823	0.960	0.819	0.905	0.041
1.0	0.833	0.969	0.824	0.911	0.122
1.5	0.848	0.979	0.860	0.852	0.103
2.0	0.811	0.951	0.772	0.888	0.232
2.5	0.724	1.000	0.724	1.000	-
3.0	0.721	0.917	0.688	0.839	0.158
4.0	0.843	0.929	0.650	0.835	0.332
5.0	0.788	0.917	0.647	0.938	0.357
6.0	0.812	0.900	0.643	0.845	0.468
All	0.825	0.965	0.806	0.916	0.147

Note. Classification performance metrics presented at individual time points as well as across all time points. NRI based on reclassification from logistic reference model; NRI incalculable under certain ROC AUC conditions.

Table 40

Evaluation of Boosted Trees Model Relative to Logistic Reference – Whole Subject Trajectories of CDR-Based Impairment

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
Accuracy	0.825	0.711	0.114	16.05%	[0.090, 0.135]
Precision	0.965	0.958	0.007	0.76%	[-0.007, 0.019]
Recall	0.806	0.661	0.146	22.03%	[0.118, 0.169]
ROC AUC	0.916	0.841	0.075	8.87%	[0.059, 0.090]

Note. Values for change in metric and percent difference relative to logistic reference model; bootstrap confidence interval corresponds to change in metric.

Figure 45

Boosted Trees Model True and Predicted CDR-Based Impairment Counts – Whole Subject Trajectories

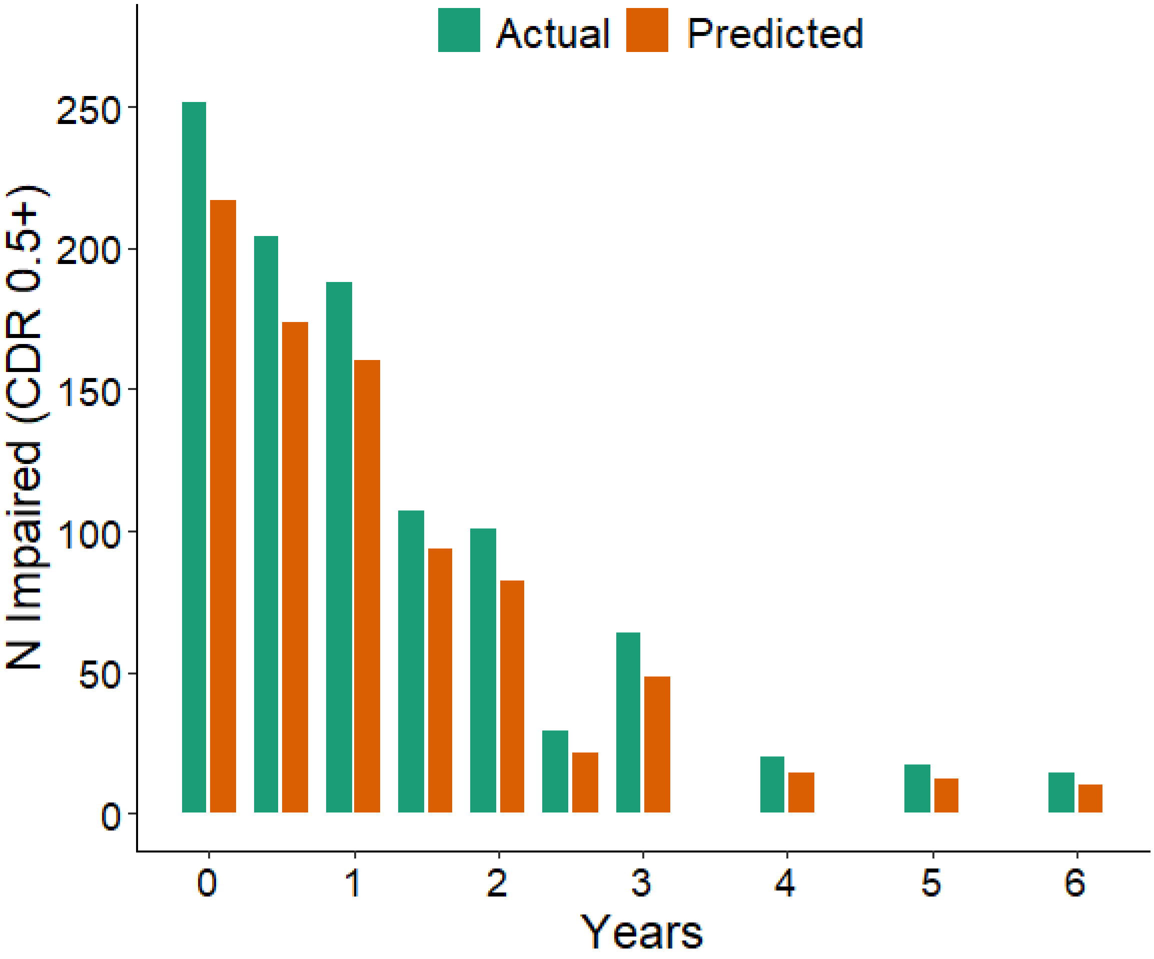
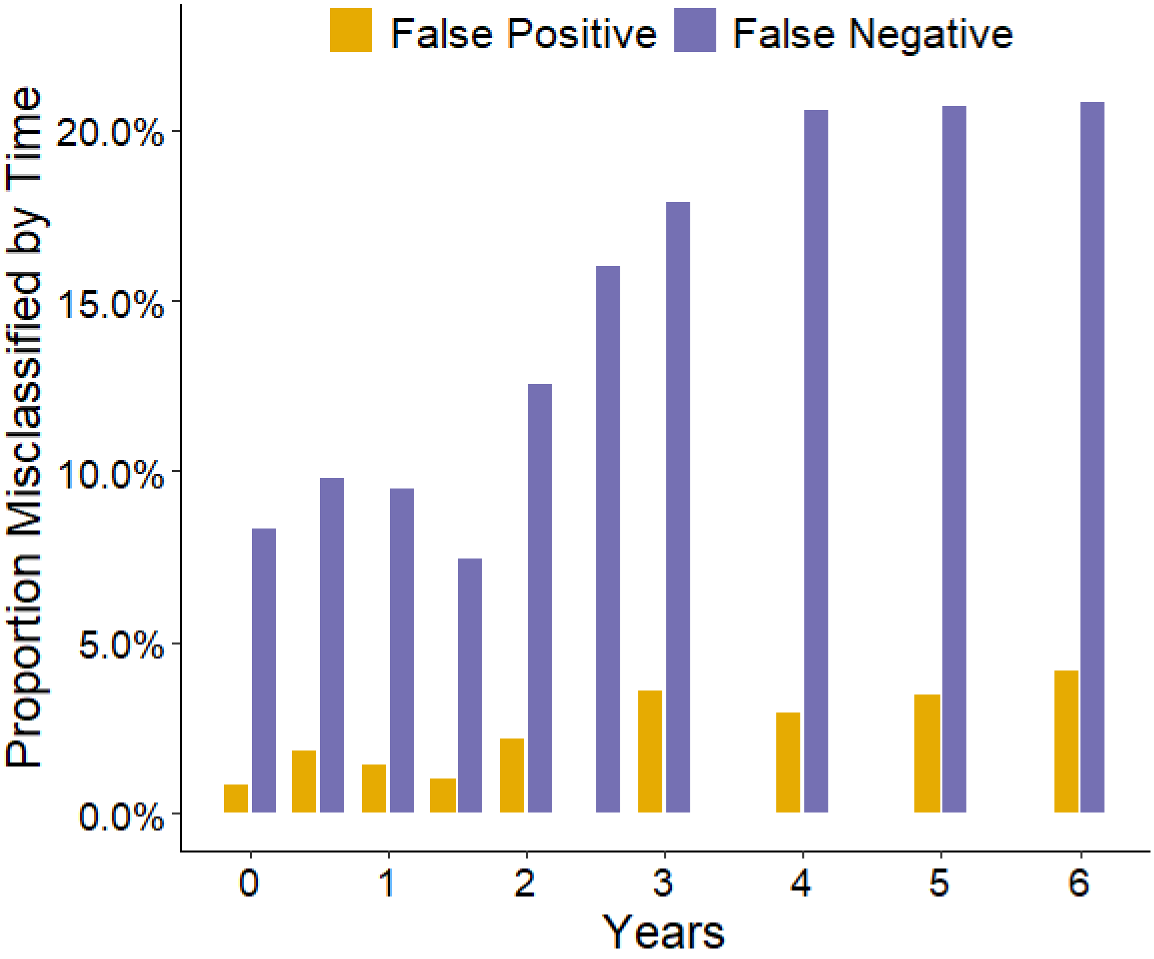


Figure 46

Boosted Trees Model Misclassification Rates for CDR-Based Impairment – Whole Subject Trajectories



Note. False positive and false negative rates marginalized by time to be relative to total counts at that timepoint.

Table 41

Boosted Trees Model Prediction Performance – Final Observation Forecasts of CDR-Based Impairment

Years	Accuracy	Precision	Recall	AUC	NRI
0.5	0.857	1.000	0.846	1.000	1.231
1.0	0.967	1.000	0.967	1.000	-
1.5	1.000	1.000	1.000	1.000	-
2.0	0.944	0.967	0.967	0.978	0.167
3.0	0.913	0.950	0.950	0.854	0.233
4.0	0.933	1.000	0.900	0.980	0.400
5.0	0.833	1.000	0.714	1.000	-
6.0	0.935	1.000	0.895	0.961	0.474
All	0.935	0.982	0.939	0.972	0.246

Note. Classification performance metrics presented at individual time points as well as across all time points. NRI based on reclassification from logistic reference model; NRI incalculable under certain ROC AUC conditions.

Table 42

Evaluation of Boosted Trees Model Relative to Logistic Reference – Final Observation Forecasts of CDR-Based Impairment

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
Accuracy	0.935	0.730	0.204	27.98%	[0.172, 0.232]
Precision	0.982	0.977	0.006	0.60%	[-0.015, 0.023]
Recall	0.939	0.694	0.244	35.15%	[0.206, 0.277]
ROC AUC	0.972	0.825	0.148	17.90%	[0.123, 0.168]

Note. Values for change in metric and percent difference relative to logistic reference model; bootstrap confidence interval corresponds to change in metric.

Figure 47

Boosted Trees Model True and Predicted CDR-Based Impairment Counts – Final Observation Forecasts

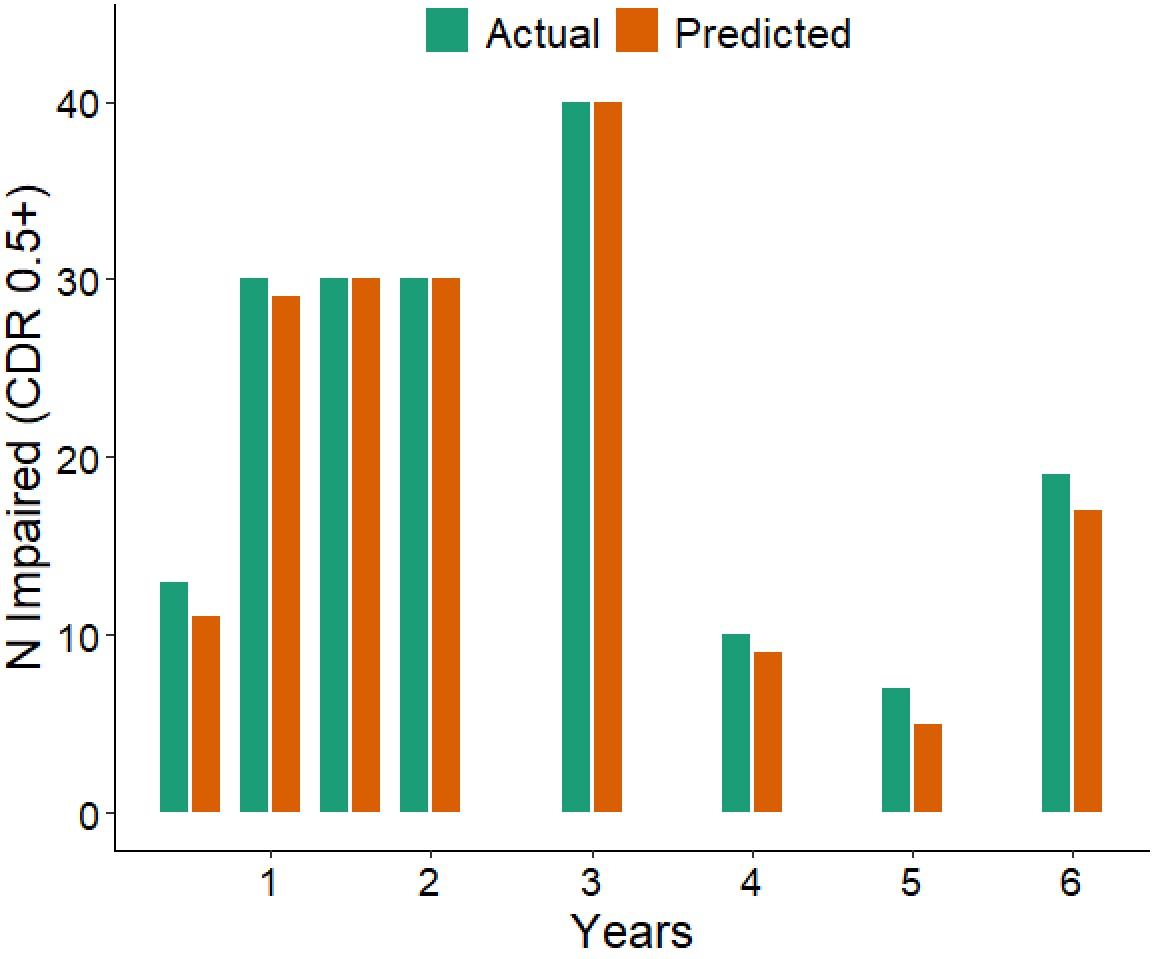
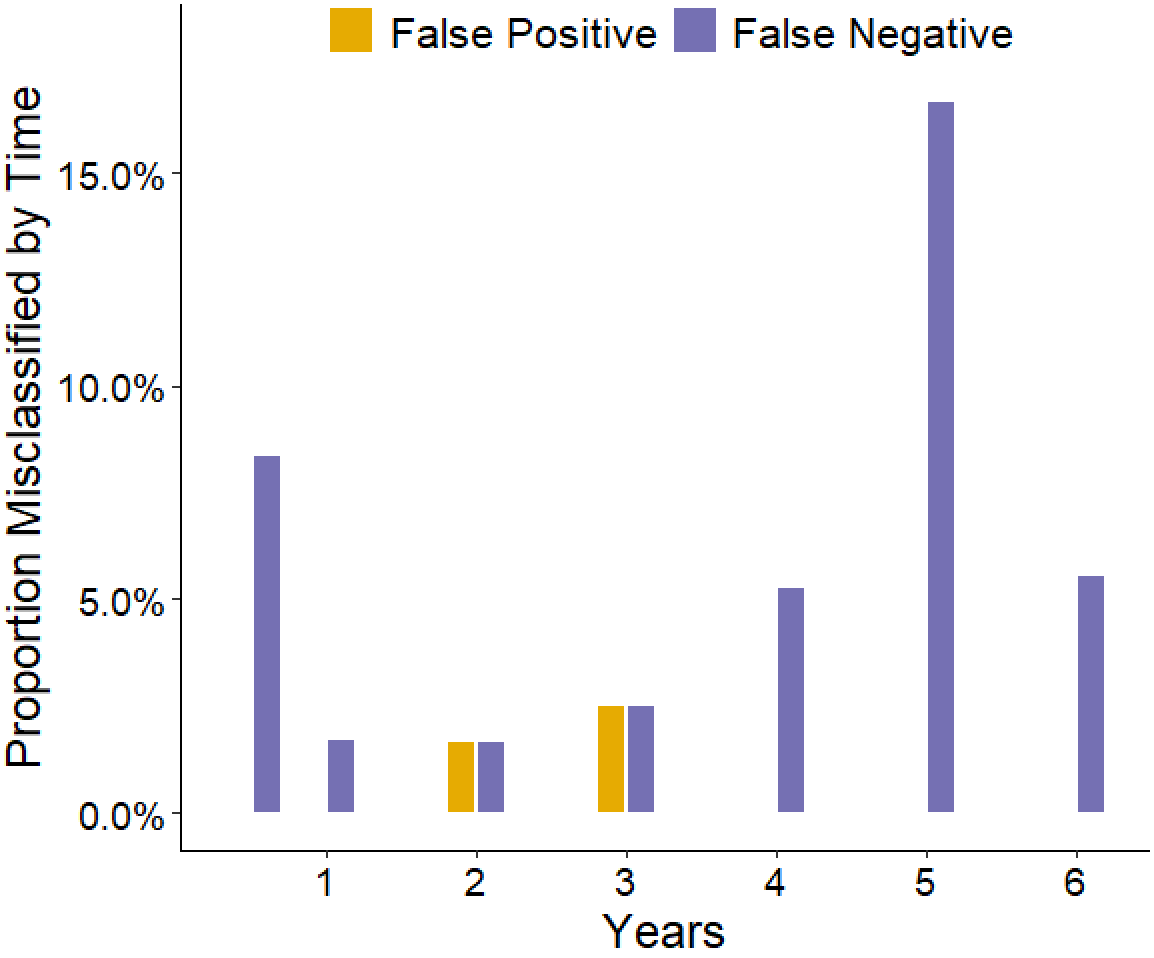


Figure 48

Boosted Trees Model Misclassification Rates for CDR-Based Impairment – Final Observation Forecasts



Note. False positive and false negative rates marginalized by time to be relative to total counts at that timepoint.

Table 43

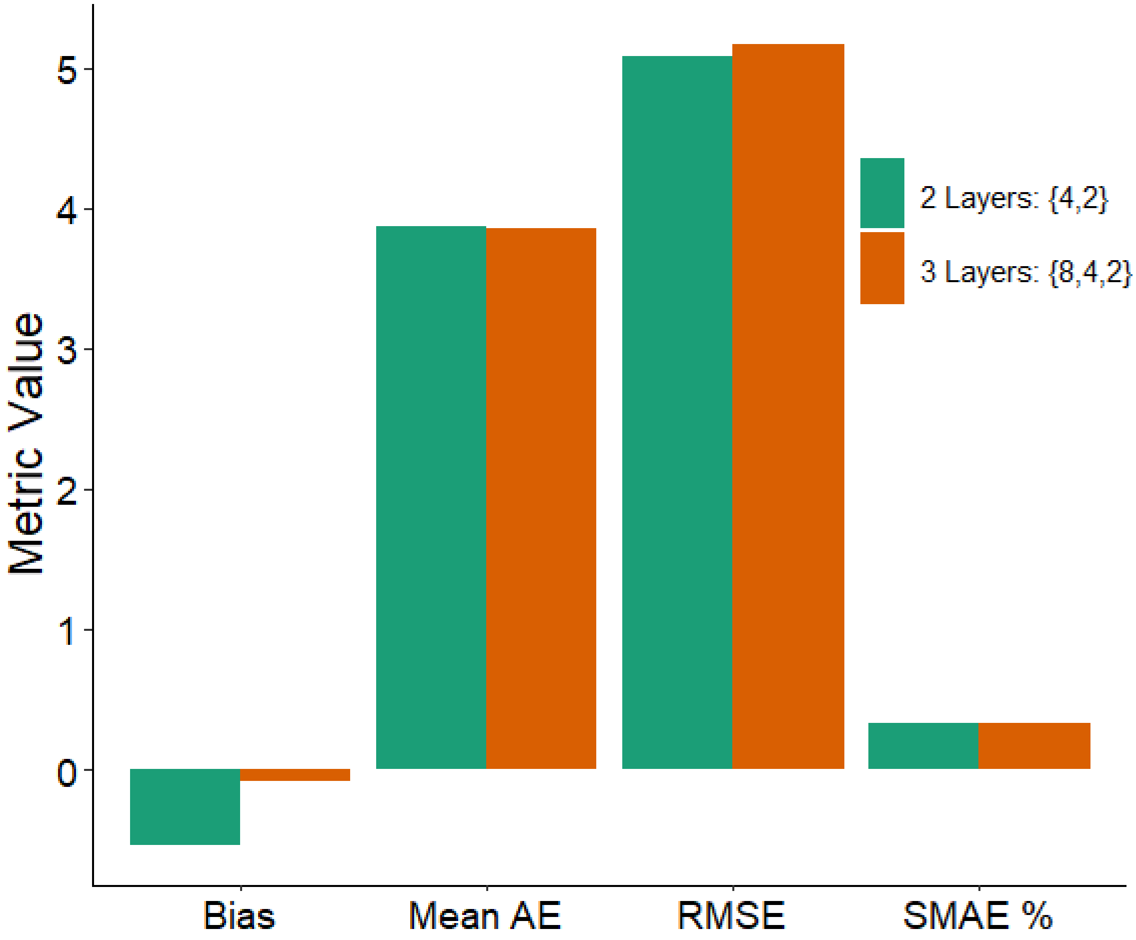
Neural Networks Performance Summary – ADAS-Cog Score

Type of prediction	Performance metric	CPath reference	FNN	1D CNN	LSTM RNN
Whole subject trajectories					
	RMSE	6.819	4.660	6.506	6.167
	Mean AE	5.283	3.560	4.673	4.423
	Bias	1.602	0.526	1.075	1.359
	AV bias	4.421	2.830	3.427	3.146
Final observation forecasts					
	RMSE	9.720	5.483	5.856	5.650
	Mean AE	6.859	4.257	3.823	3.727
	Bias	-1.360	0.431	-1.068	-1.109
	AV bias	4.799	3.437	2.280	2.438

Note. Regression performance metrics summarized across all time points with CPath model shown for reference.

Figure 49

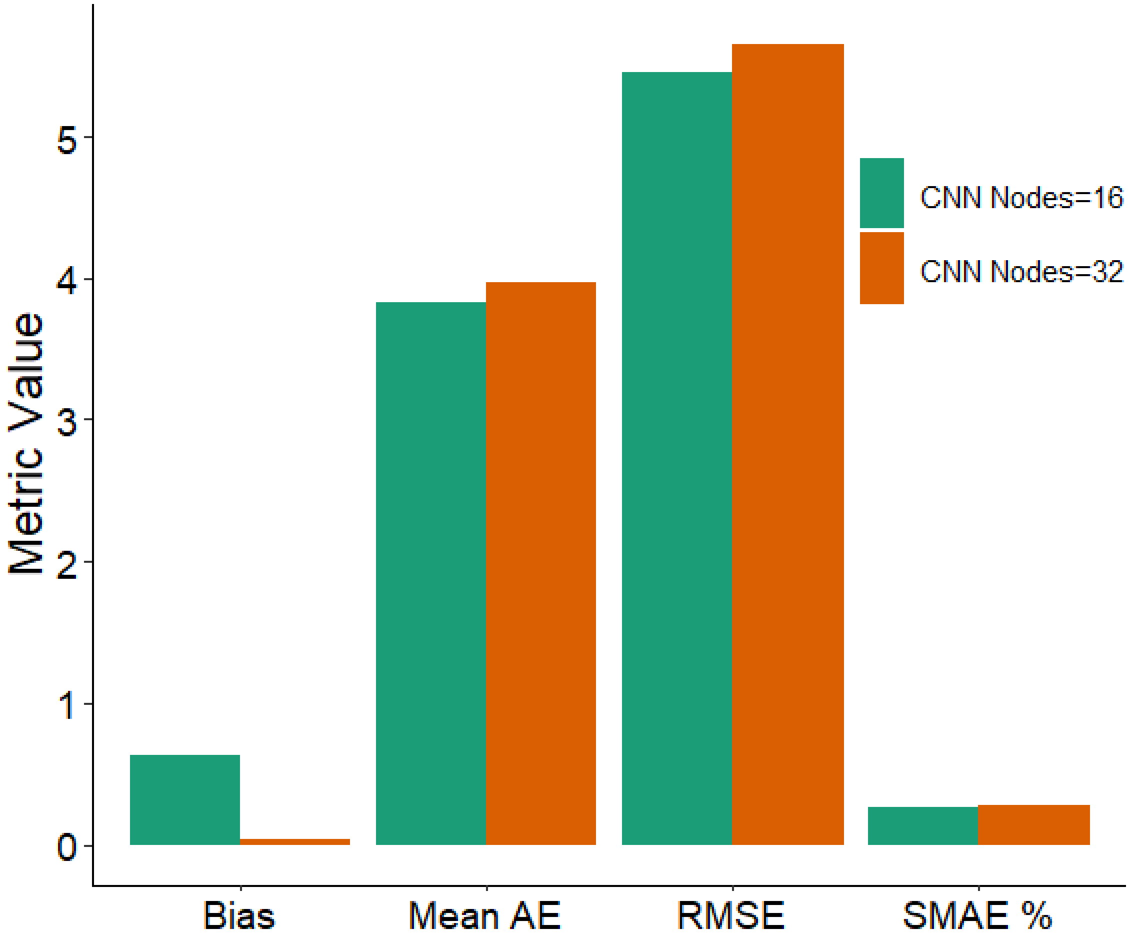
Feed-Forward Neural Network Hyperparameter Tunings – ADAS-Cog Score



Note. Hyperparameter tunings provided from cross-fold validations on whole subject trajectories.

Figure 50

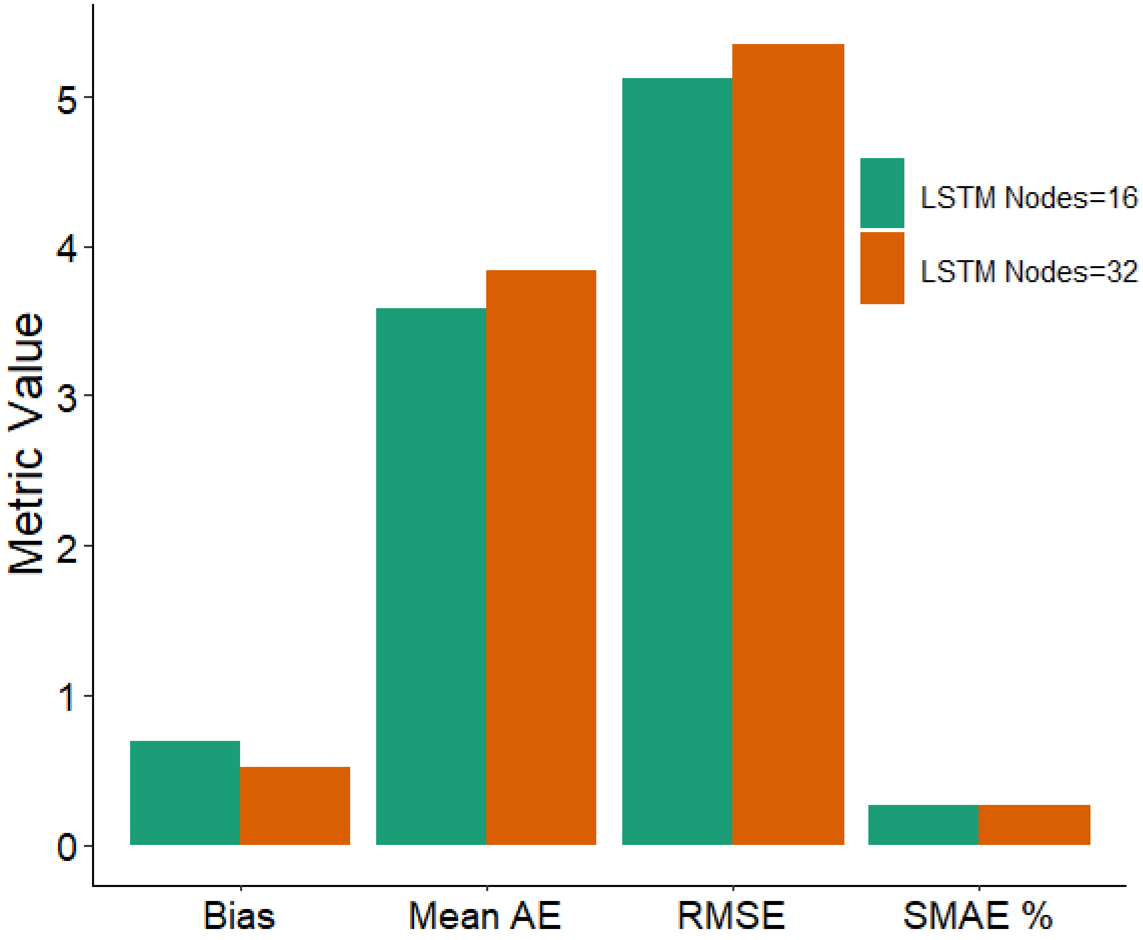
1D Convolutional Neural Network Hyperparameter Tunings – ADAS-Cog Score



Note. Hyperparameter tunings provided from cross-fold validations on whole subject trajectories.

Figure 51

LSTM Recurrent Neural Network Hyperparameter Tunings – ADAS-Cog Score



Note. Hyperparameter tunings provided from cross-fold validations on whole subject trajectories.

Table 44

Top Hyperparameters for Neural Networks – ADAS-Cog Score

Neural network	Performance metric	Best value	Hyperparameter set
Feed-forward neural network			
	RMSE	5.087	2 layers: {4,2}
	Mean AE	3.857	3 layers: {8,4,2}
	SMAE %	0.328	3 layers: {8,4,2}
	Bias	-0.092	3 layers: {8,4,2}
1D convolutional neural network			
	RMSE	5.458	CNN nodes=16
	Mean AE	3.831	CNN nodes=16
	SMAE %	0.267	CNN nodes=16
	Bias	-0.030	CNN nodes=32
LSTM recurrent neural network			
	RMSE	5.127	LSTM nodes=16
	Mean AE	3.581	LSTM nodes=16
	SMAE %	0.259	LSTM nodes=16
	Bias	-0.519	LSTM nodes=32

Note. For each neural network, the best performing hyperparameter set for each regression metric is displayed along with the corresponding value.

Table 45

Feed-Forward Neural Network Prediction Performance – Whole Subject Trajectories of ADAS-Cog Score

Years	RMSE	MAE	Bias	AV Bias
0.0	4.709	3.502	0.545	2.648
0.25	5.477	4.344	0.345	3.233
0.5	4.356	3.350	0.971	2.617
0.75	5.008	4.137	-0.088	3.336
1.0	4.868	3.700	0.603	2.937
1.25	5.781	4.467	0.444	3.682
1.5	4.507	3.365	0.490	2.747
2.0	4.731	3.684	0.571	3.299
2.5	4.465	3.429	-1.707	2.473
3.0	4.552	3.591	0.317	2.947
4.0	3.711	2.929	0.761	2.696
5.0	3.265	2.386	0.426	1.854
6.0	4.060	3.284	0.312	2.827
All	4.660	3.560	0.526	2.830

Note. Regression performance metrics presented at individual time points as well as across all time points.

Table 46

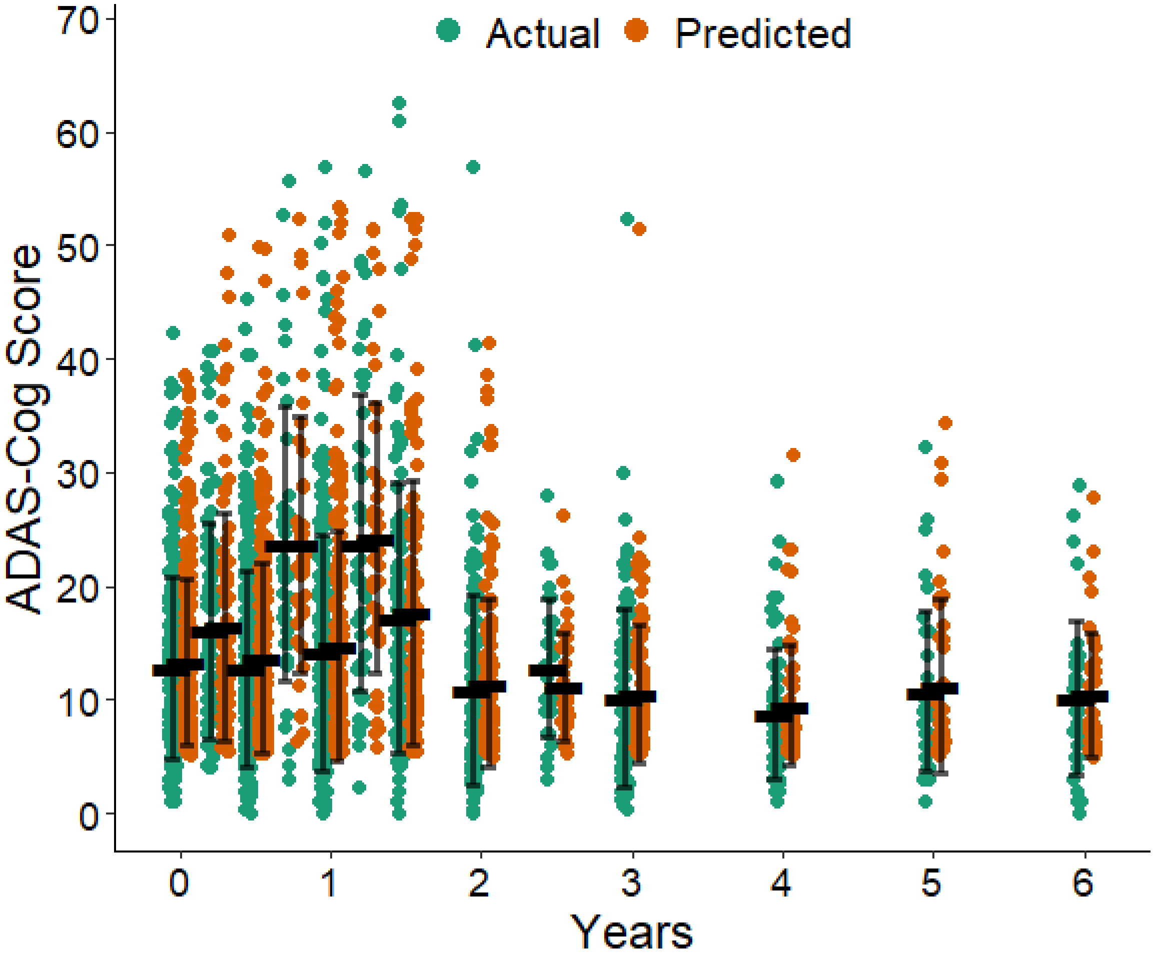
Evaluation of Feed-Forward Neural Network Relative to CPath Reference – Whole Subject Trajectories of ADAS-Cog Score

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
RMSE	4.660	6.819	-2.160	-31.67%	[-2.351, -1.972]
Mean AE	3.560	5.283	-1.723	-32.61%	[-1.876, -1.579]
Bias	0.526	1.602	-1.076	-67.18%	[-1.297, -0.853]
AV bias	2.830	4.421	-1.591	-35.99%	[-1.733, -1.429]

Note. Values for change in metric and percent difference relative to CPath reference model; bootstrap confidence interval corresponds to change in metric.

Figure 52

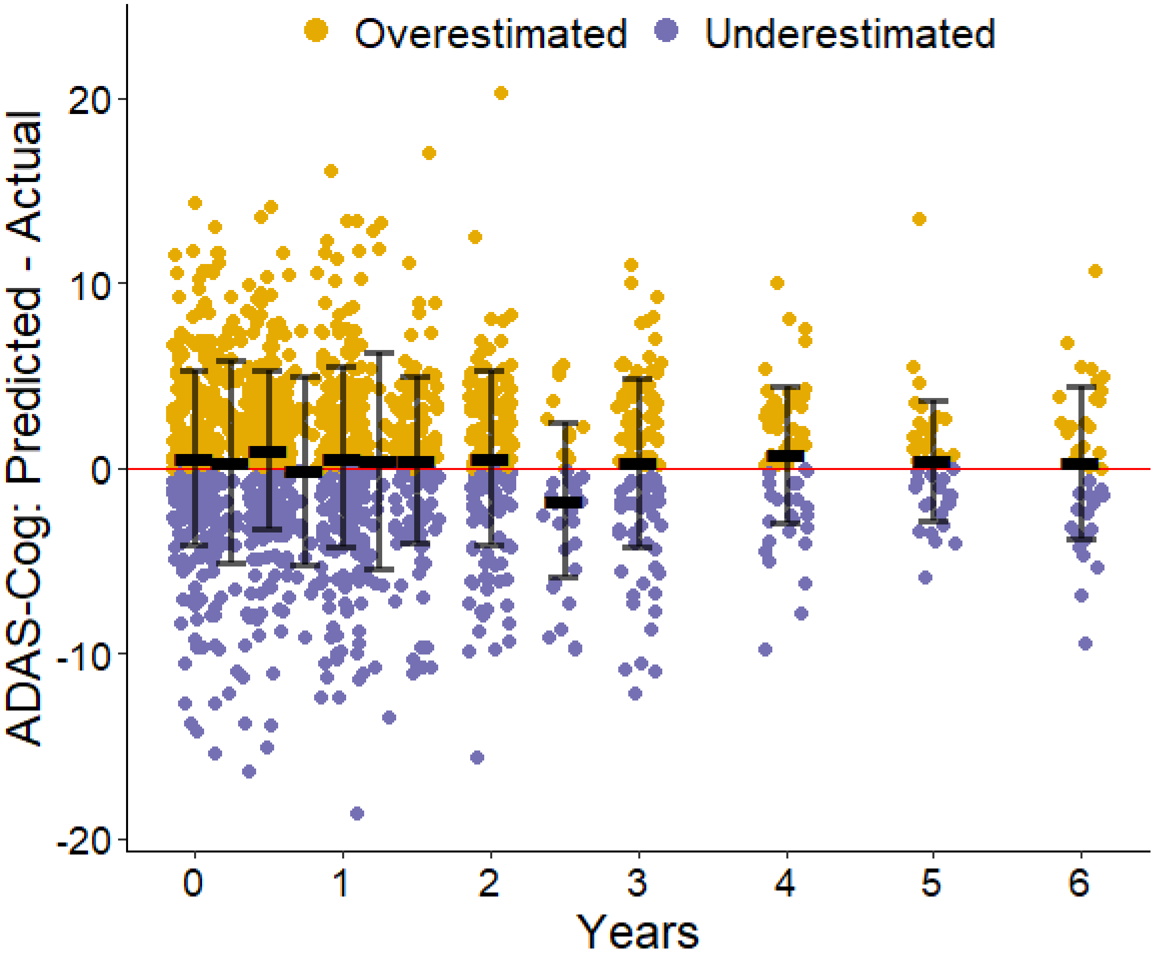
Feed-Forward Neural Network True and Predicted ADAS-Cog Scores – Whole Subject Trajectories



Note. Error bars centered at mean with standard deviation ranges.

Figure 53

Feed-Forward Neural Network Prediction Discrepancies for ADAS-Cog Scores – Whole Subject Trajectories



Note. Error bars centered at mean with standard deviation ranges.

Table 47

Feed-Forward Neural Network Prediction Performance – Final Observation Forecasts of ADAS-Cog Score

Years	RMSE	MAE	Bias	AV Bias
0.25	5.017	4.021	0.958	3.075
0.5	3.874	3.153	0.400	2.633
1.0	5.363	4.146	0.810	3.611
1.25	6.575	4.725	2.563	3.721
1.5	7.189	6.127	3.179	6.241
2.0	6.671	5.097	0.035	3.883
3.0	4.817	3.881	-0.082	3.276
4.0	4.768	3.638	0.230	2.597
5.0	2.553	1.976	-0.679	1.530
6.0	5.004	4.022	-0.912	3.142
All	5.483	4.257	0.431	3.437

Note. Regression performance metrics presented at individual time points as well as across all time points.

Table 48

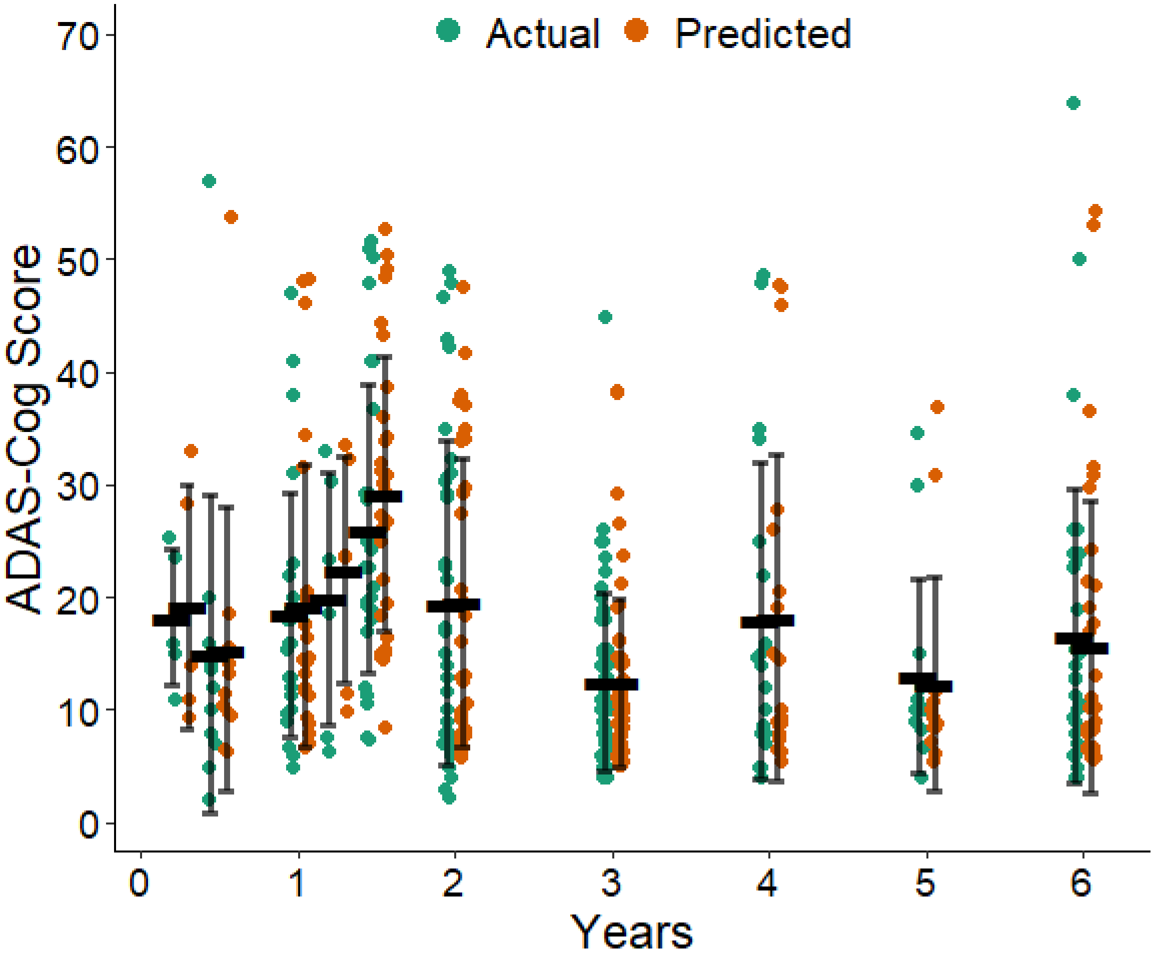
Evaluation of Feed-Forward Neural Network Relative to CPath Reference – Final Observation Forecasts of ADAS-Cog Score

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
RMSE	5.483	9.720	-4.237	-43.59%	[-4.814, -3.699]
Mean AE	4.257	6.859	-2.602	-37.94%	[-3.051, -2.182]
Bias	0.431	-1.360	1.792	-68.31%	[1.096, 2.584]
AV bias	3.437	4.799	-1.362	-28.38%	[-1.962, -0.778]

Note. Values for change in metric and percent difference relative to CPath reference model; bootstrap confidence interval corresponds to change in metric.

Figure 54

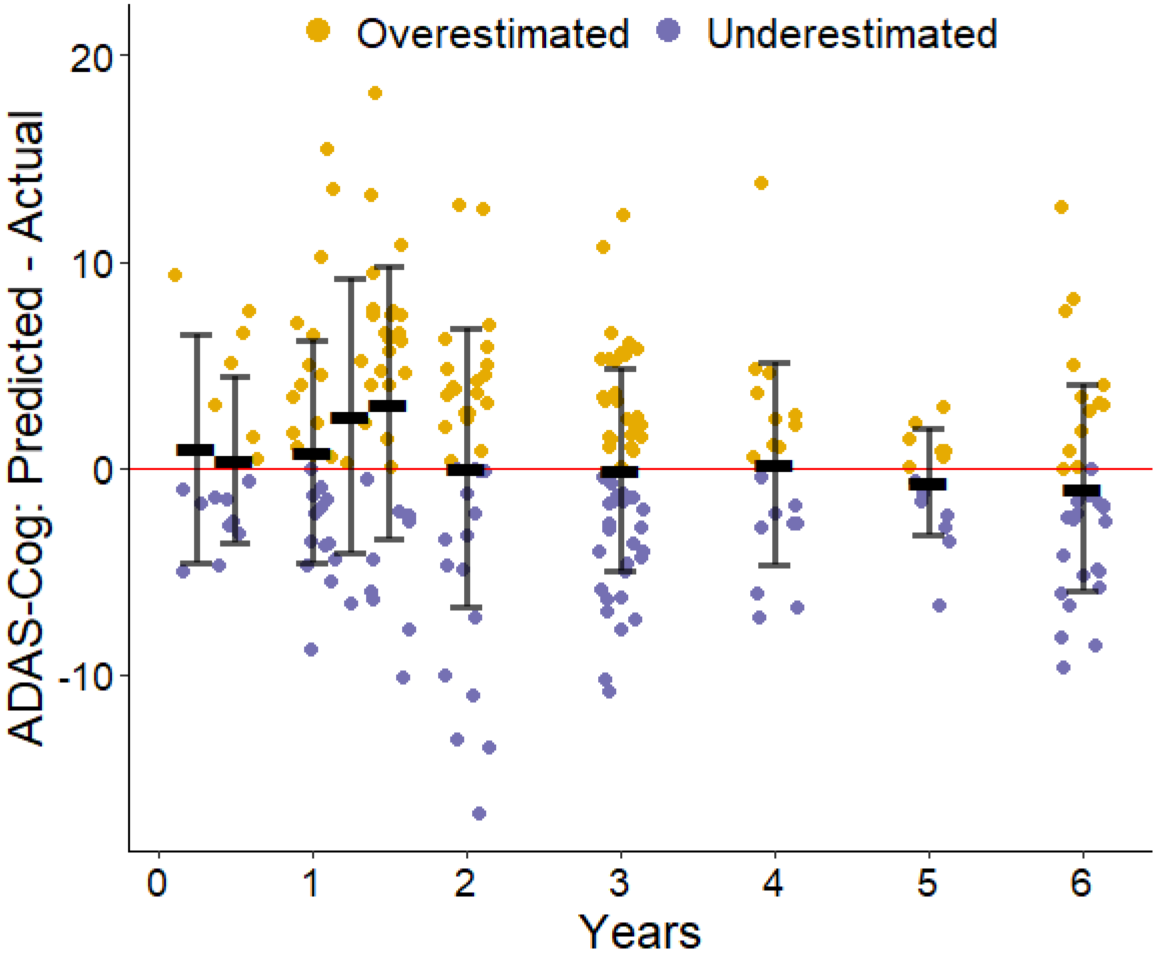
Feed-Forward Neural Network True and Predicted ADAS-Cog Scores – Final Observation Forecasts



Note. Error bars centered at mean with standard deviation ranges.

Figure 55

Feed-Forward Neural Network Prediction Discrepancies for ADAS-Cog Scores – Final Observation Forecasts



Note. Error bars centered at mean with standard deviation ranges.

Table 49

1D Convolutional Neural Network Prediction Performance – Whole Subject Trajectories of ADAS-Cog Score

Years	RMSE	MAE	Bias	AV Bias
0.0	4.597	3.463	0.769	2.520
0.25	9.456	7.709	0.111	6.869
0.5	5.118	3.976	1.078	3.043
0.75	8.418	6.389	-1.997	4.792
1.0	6.292	4.549	0.407	3.242
1.25	9.407	7.547	0.743	5.704
1.5	7.397	5.312	0.916	3.790
2.0	5.820	4.284	1.087	3.294
2.5	6.148	4.712	2.644	3.372
3.0	6.440	4.764	2.362	4.087
4.0	5.905	3.755	1.998	2.425
5.0	9.299	5.600	3.631	2.887
6.0	9.135	5.704	4.614	3.101
All	6.506	4.673	1.075	3.427

Note. Regression performance metrics presented at individual time points as well as across all time points.

Table 50

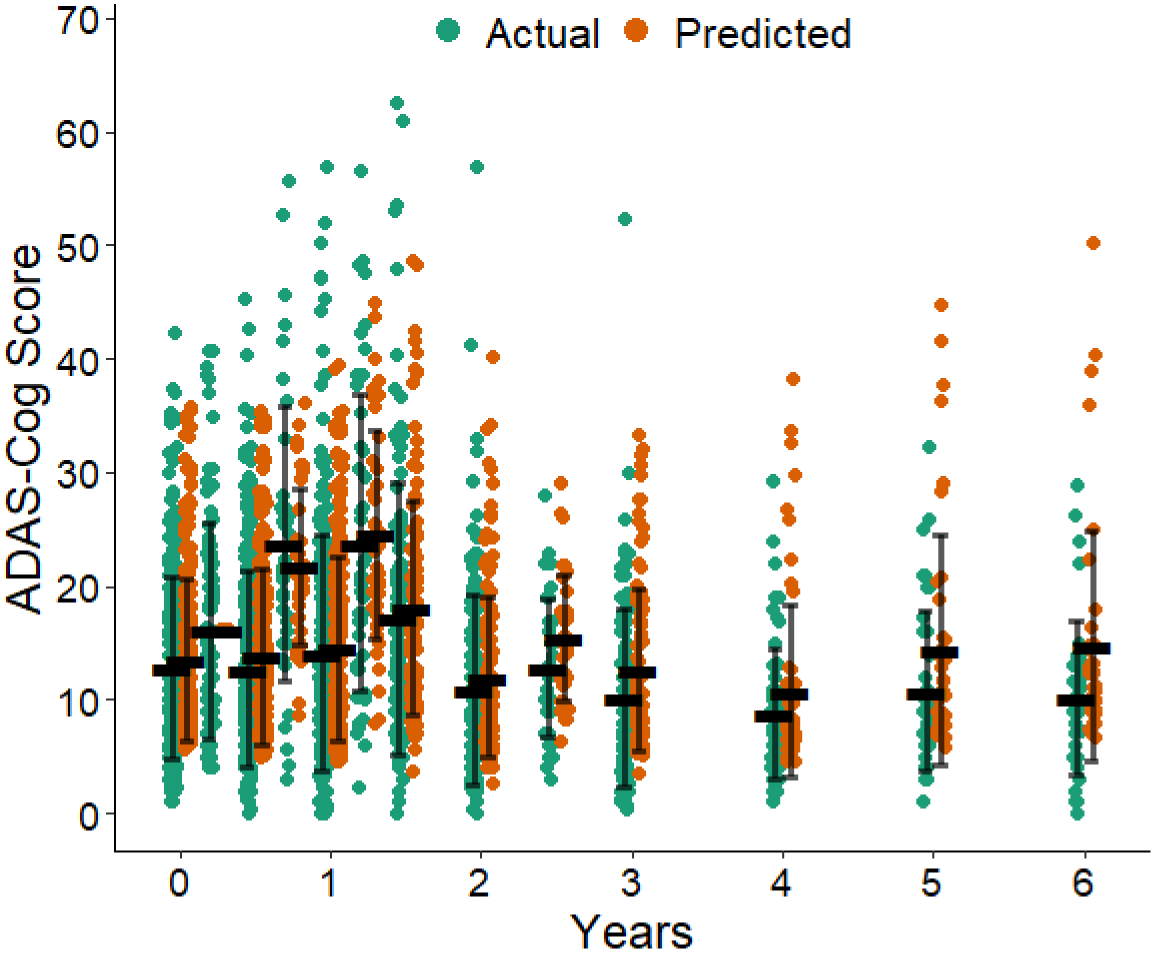
Evaluation of 1D Convolutional Neural Network Relative to CPath Reference – Whole Subject Trajectories of ADAS-Cog Score

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
RMSE	6.506	6.819	-0.313	-4.59%	[-0.705, 0.050]
Mean AE	4.673	5.283	-0.610	-11.54%	[-0.845, -0.397]
Bias	1.075	1.602	-0.527	-32.88%	[-0.885, -0.219]
AV bias	3.427	4.421	-0.994	-22.48%	[-1.230, -0.771]

Note. Values for change in metric and percent difference relative to CPath reference model; bootstrap confidence interval corresponds to change in metric.

Figure 56

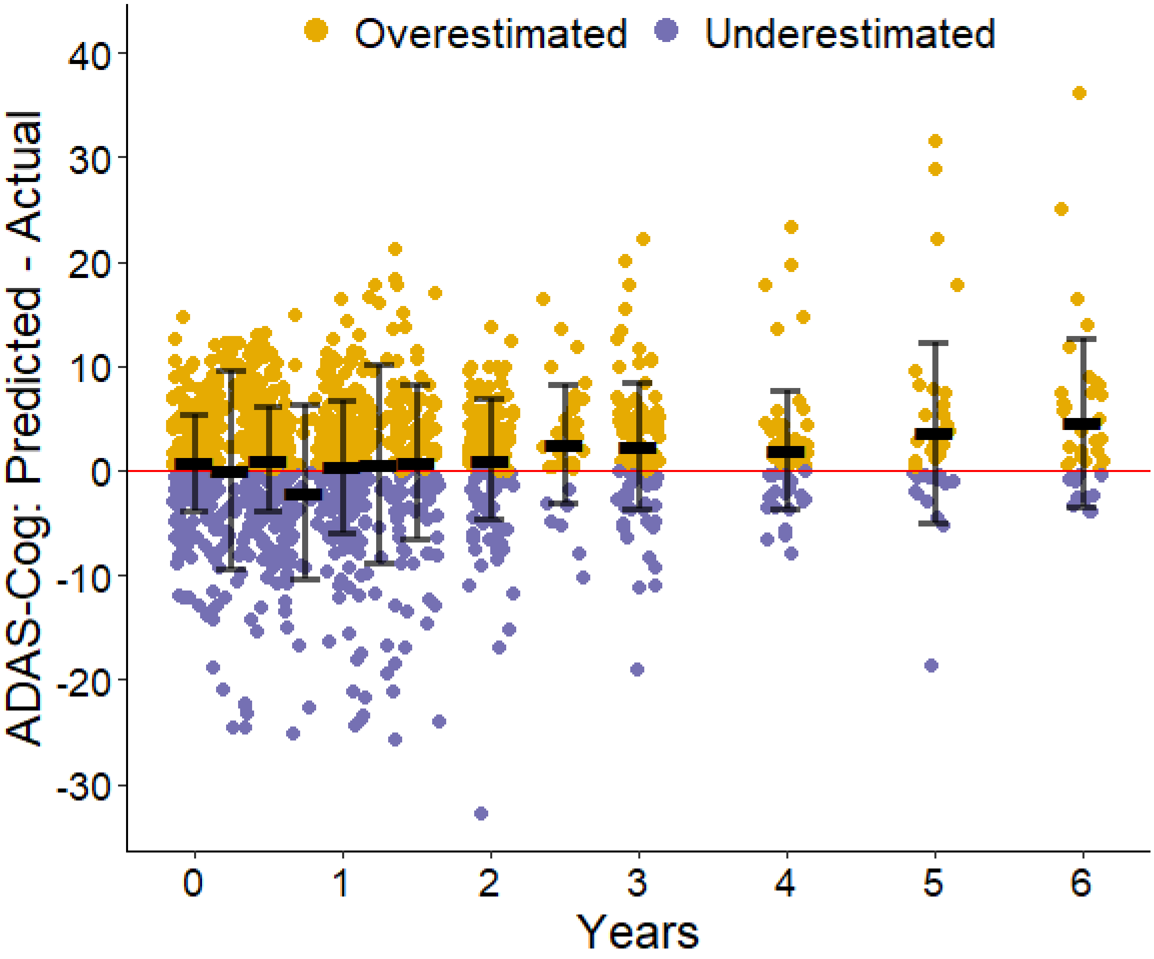
1D Convolutional Neural Network True and Predicted ADAS-Cog Scores – Whole Subject Trajectories



Note. Error bars centered at mean with standard deviation ranges.

Figure 57

1D Convolutional Neural Network Prediction Discrepancies for ADAS-Cog Scores – Whole Subject Trajectories



Note. Error bars centered at mean with standard deviation ranges.

Table 51

1D Convolutional Neural Network Prediction Performance – Final Observation Forecasts of ADAS-Cog Score

Years	RMSE	MAE	Bias	AV Bias
0.25	5.791	4.640	-2.001	5.199
0.5	7.186	4.663	-2.023	2.662
1.0	5.896	4.042	-1.393	3.511
1.25	5.782	4.777	1.534	4.384
1.5	7.752	5.155	0.156	3.219
2.0	4.857	3.452	-0.933	2.617
2.5	2.646	2.646	2.646	2.646
3.0	3.909	2.671	-0.671	1.803
4.0	8.314	5.065	-2.712	2.662
5.0	4.697	3.093	-0.715	1.360
6.0	5.633	3.732	-1.894	1.983
All	5.856	3.823	-1.068	2.280

Note. Regression performance metrics presented at individual time points as well as across all time points.

Table 52

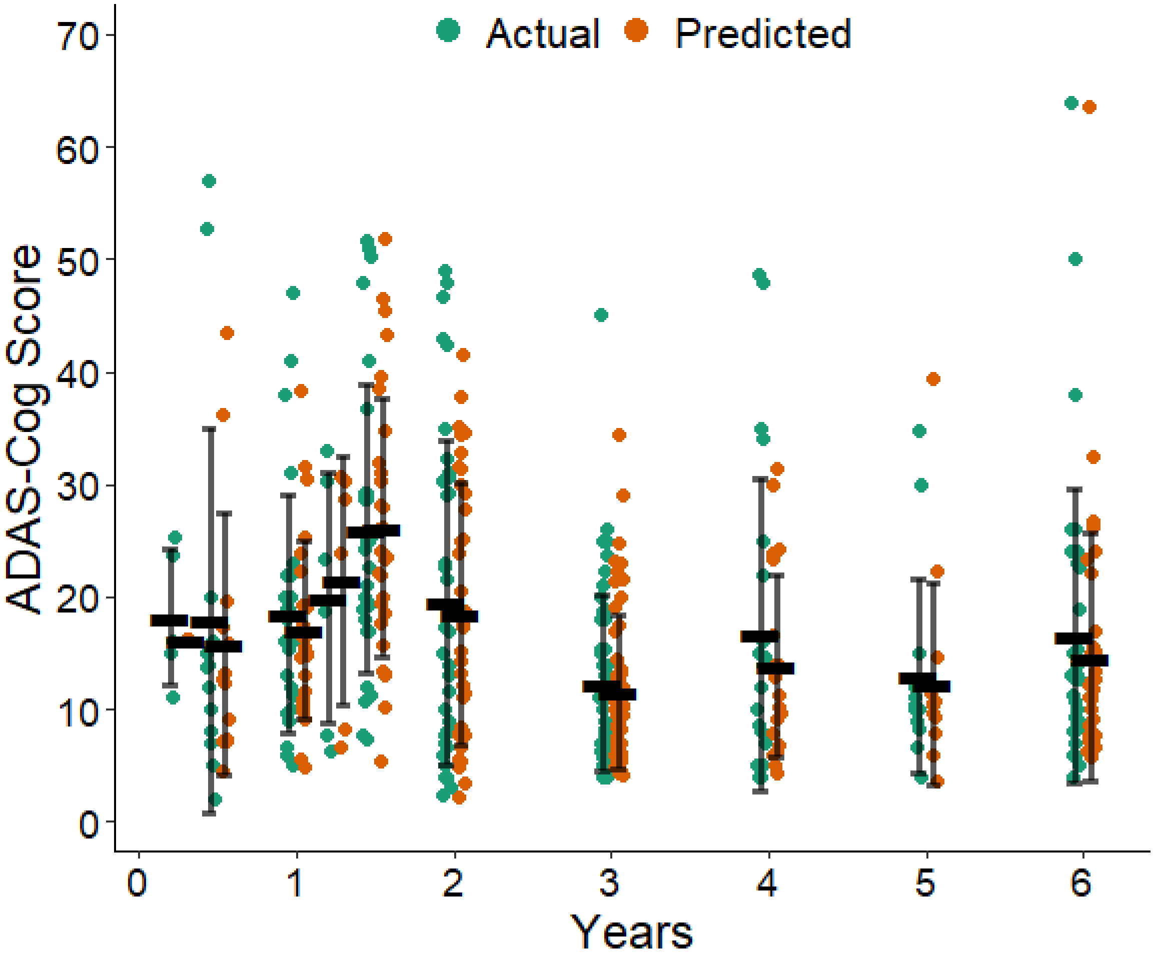
Evaluation of 1D Convolutional Neural Network Relative to CPath Reference – Final Observation Forecasts of ADAS-Cog Score

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
RMSE	5.856	9.720	-3.864	-39.76%	[-4.822, -2.844]
Mean AE	3.823	6.859	-3.035	-44.26%	[-3.528, -2.449]
Bias	-1.068	-1.360	0.292	-21.47%	[-0.412, 1.062]
AV bias	2.280	4.799	-2.519	-52.49%	[-2.857, -2.011]

Note. Values for change in metric and percent difference relative to CPath reference model; bootstrap confidence interval corresponds to change in metric.

Figure 58

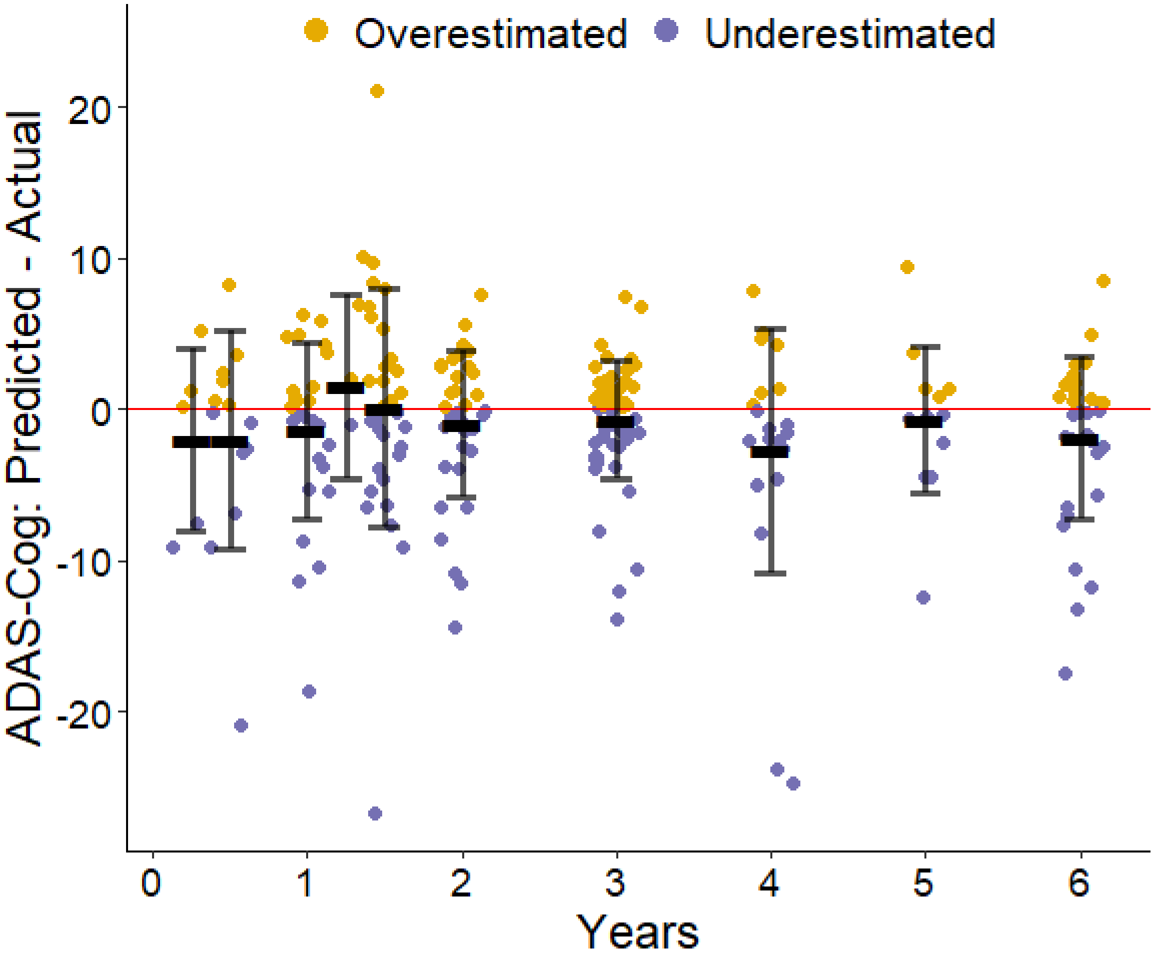
1D Convolutional Neural Network True and Predicted ADAS-Cog Scores – Final Observation Forecasts



Note. Error bars centered at mean with standard deviation ranges.

Figure 59

1D Convolutional Neural Network Prediction Discrepancies for ADAS-Cog Scores – Final Observation Forecasts



Note. Error bars centered at mean with standard deviation ranges.

Table 53

LSTM Recurrent Neural Network Prediction Performance – Whole Subject Trajectories of ADAS-Cog Score

Years	RMSE	MAE	Bias	AV Bias
0.0	4.561	3.442	0.707	2.490
0.25	5.985	4.810	0.709	3.982
0.5	5.240	3.966	1.298	2.975
0.75	8.993	6.459	3.051	3.788
1.0	6.439	4.591	0.686	3.162
1.25	9.410	7.334	3.439	5.512
1.5	7.345	5.368	1.776	3.461
2.0	5.500	4.057	0.867	3.170
2.5	5.018	3.616	0.065	2.421
3.0	5.987	4.370	1.920	3.220
4.0	5.727	3.874	2.450	2.560
5.0	9.000	5.945	3.281	3.682
6.0	8.435	5.426	4.274	3.797
All	6.167	4.423	1.359	3.146

Note. Regression performance metrics presented at individual time points as well as across all time points.

Table 54

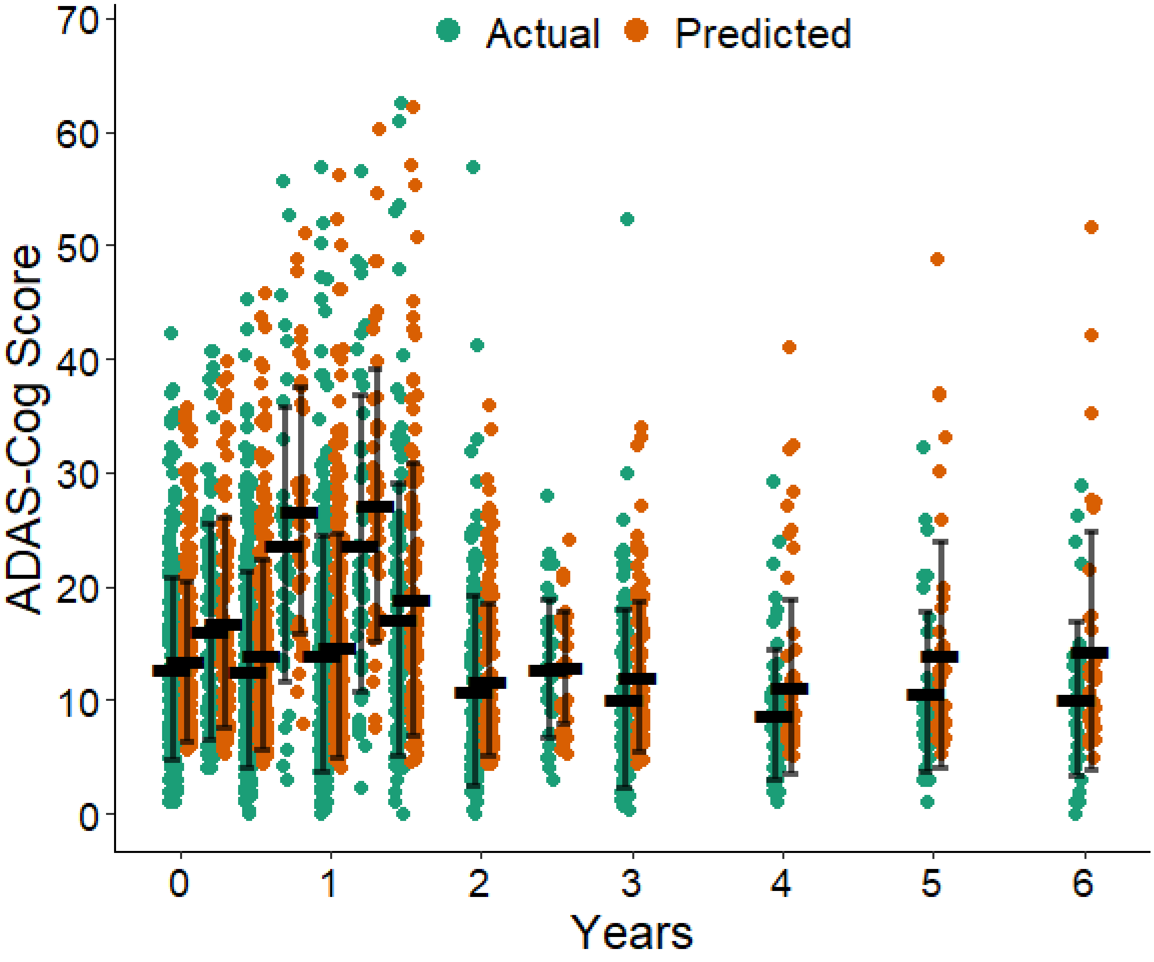
Evaluation of LSTM Recurrent Neural Network Relative to CPath Reference – Whole Subject Trajectories of ADAS-Cog Score

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
RMSE	6.167	6.819	-0.652	-9.56%	[-0.996, -0.260]
Mean AE	4.423	5.283	-0.861	-16.29%	[-1.067, -0.629]
Bias	1.359	1.602	-0.244	-15.20%	[-0.532, 0.089]
AV bias	3.146	4.421	-1.275	-28.84%	[-1.446, -1.092]

Note. Values for change in metric and percent difference relative to CPath reference model; bootstrap confidence interval corresponds to change in metric.

Figure 60

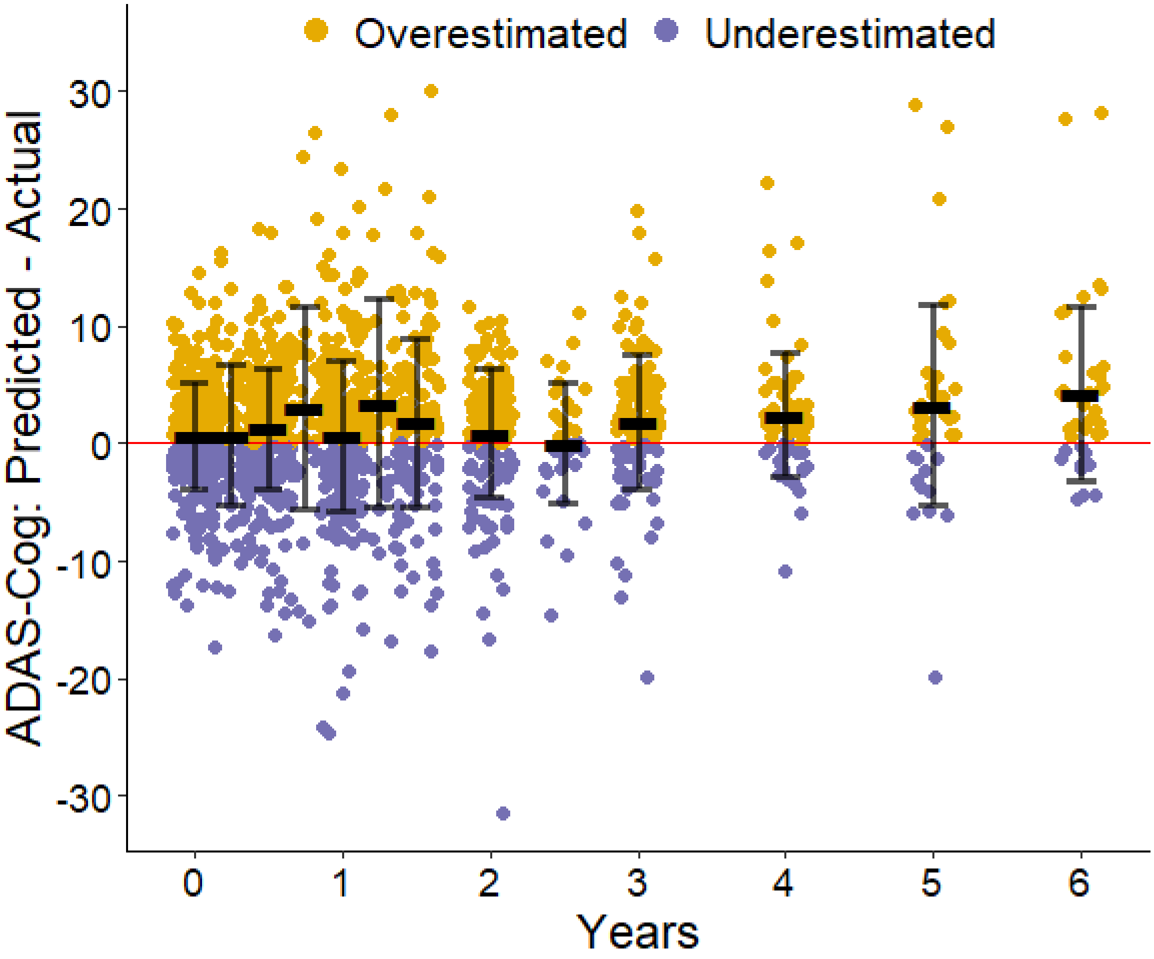
LSTM Recurrent Neural Network True and Predicted ADAS-Cog Scores – Whole Subject Trajectories



Note. Error bars centered at mean with standard deviation ranges.

Figure 61

LSTM Recurrent Neural Network Prediction Discrepancies for ADAS-Cog Scores – Whole Subject Trajectories



Note. Error bars centered at mean with standard deviation ranges.

Table 55

LSTM Recurrent Neural Network Prediction Performance – Final Observation Forecasts of ADAS-Cog Score

Years	RMSE	MAE	Bias	AV Bias
0.25	3.315	3.070	-1.764	3.584
0.5	6.723	4.546	-1.641	2.956
1.0	5.481	3.950	-1.020	2.971
1.25	6.605	5.394	1.805	5.299
1.5	7.470	5.158	0.759	3.924
2.0	5.597	3.561	-1.712	1.982
2.5	4.623	4.623	4.623	4.623
3.0	3.812	2.565	-0.732	1.777
4.0	8.071	4.369	-2.978	1.955
5.0	3.810	2.934	-0.638	2.389
6.0	4.942	3.774	-2.230	2.824
All	5.650	3.727	-1.109	2.438

Note. Regression performance metrics presented at individual time points as well as across all time points.

Table 56

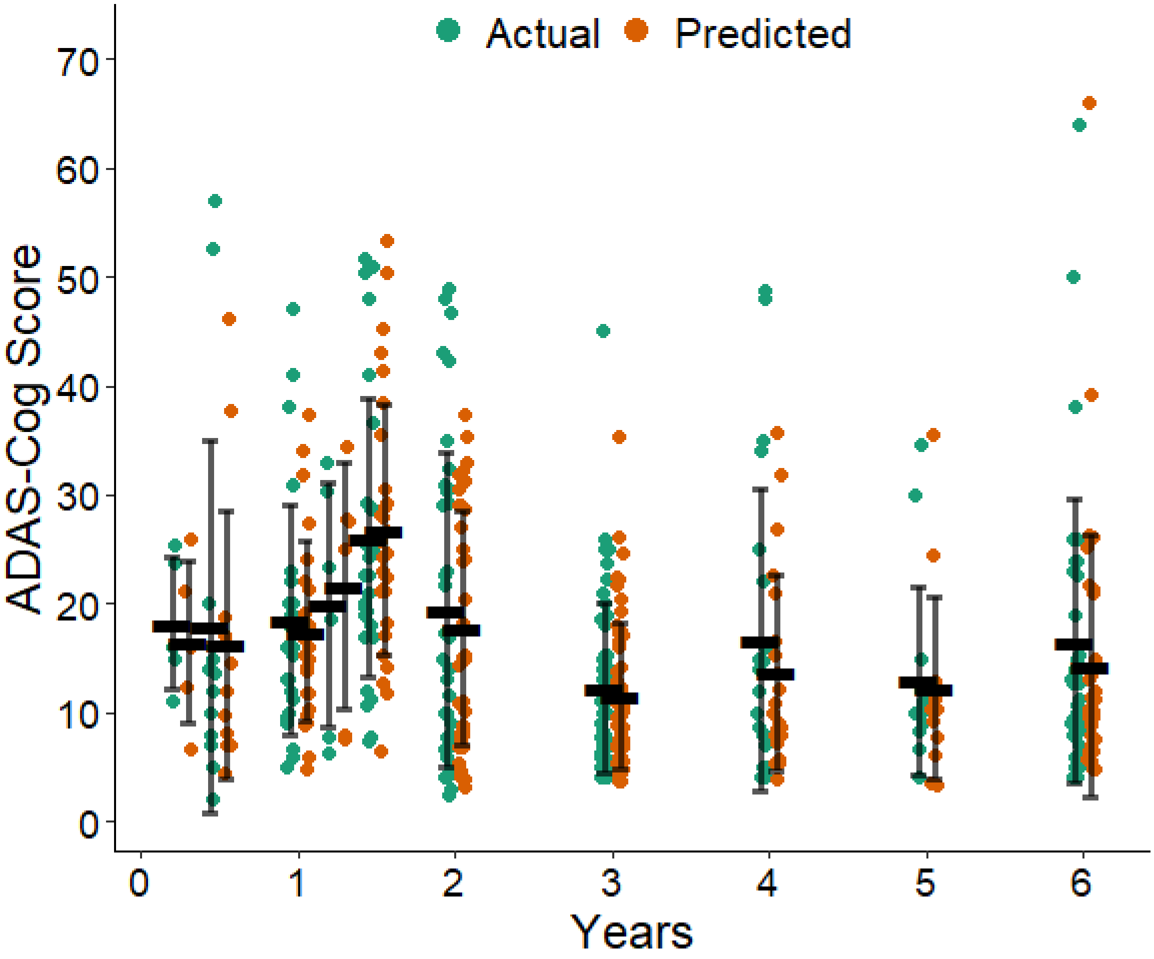
Evaluation of LSTM Recurrent Neural Network Relative to CPath Reference – Final Observation Forecasts of ADAS-Cog Score

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
RMSE	5.650	9.720	-4.070	-41.88%	[-5.099, -3.156]
Mean AE	3.727	6.859	-3.132	-45.67%	[-3.681, -2.578]
Bias	-1.109	-1.360	0.251	-18.48%	[-0.432, 0.991]
AV bias	2.438	4.799	-2.361	-49.20%	[-2.830, -1.990]

Note. Values for change in metric and percent difference relative to CPath reference model; bootstrap confidence interval corresponds to change in metric.

Figure 62

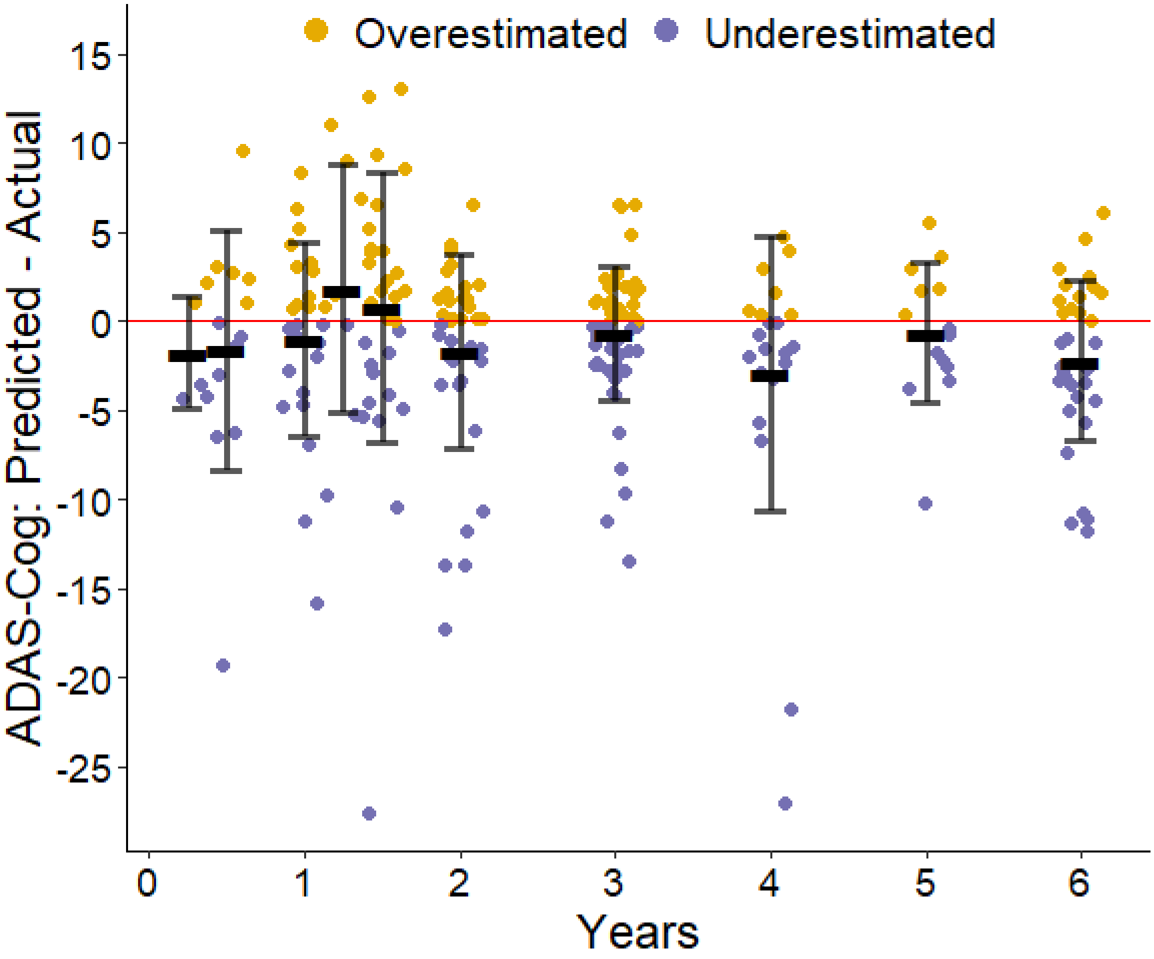
LSTM Recurrent Neural Network True and Predicted ADAS-Cog Scores – Final Observation Forecasts



Note. Error bars centered at mean with standard deviation ranges.

Figure 63

LSTM Recurrent Neural Network Prediction Discrepancies for ADAS-Cog Scores – Final Observation Forecasts



Note. Error bars centered at mean with standard deviation ranges.

Table 57

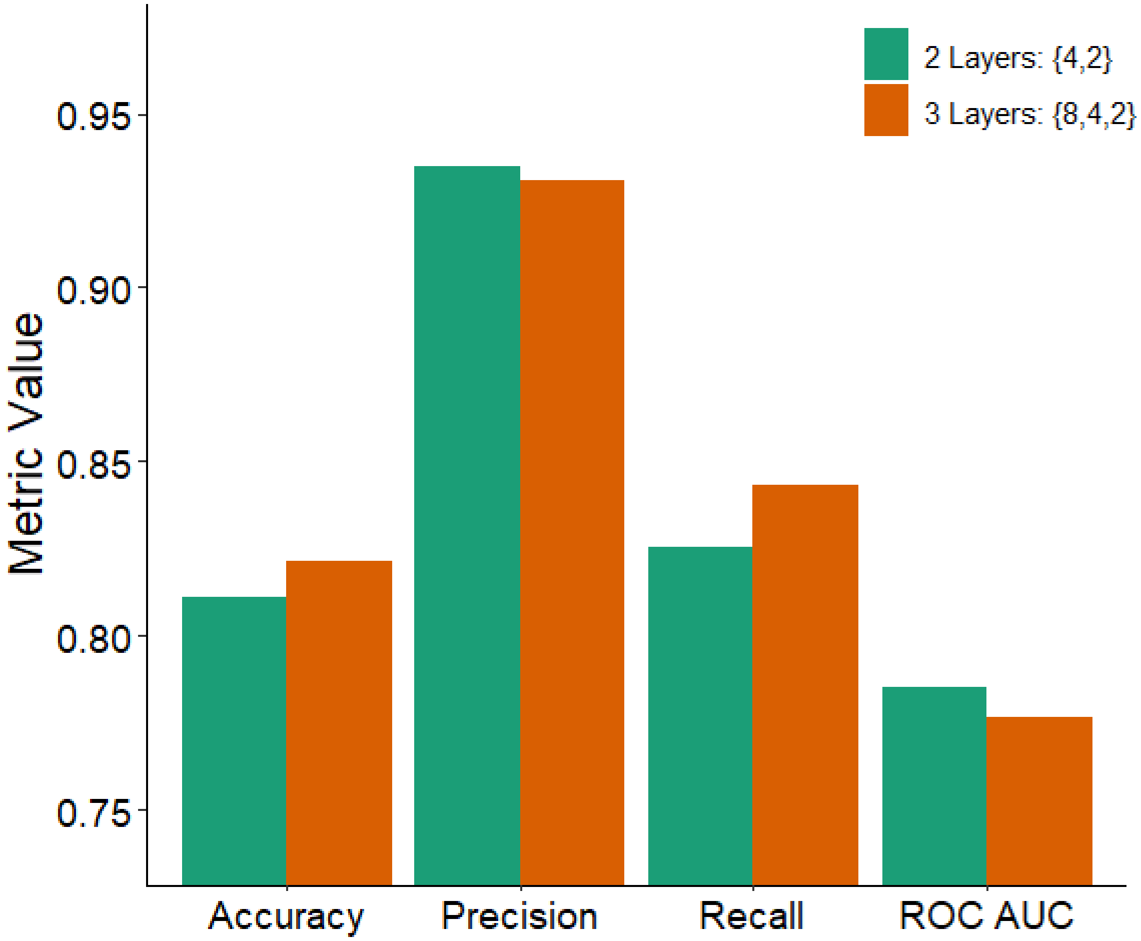
Neural Networks Performance Summary – CDR-Based Impairment

Type of prediction	Performance metric	Logistic reference	FNN	1D CNN	LSTM RNN
Whole subject trajectories					
	Accuracy	0.711	0.800	0.836	0.835
	Precision	0.958	0.972	0.963	0.958
	Recall	0.661	0.768	0.822	0.824
	ROC AUC	0.841	0.908	0.855	0.848
Final observation forecasts					
	Accuracy	0.730	0.832	0.949	0.972
	Precision	0.977	0.980	0.994	0.983
	Recall	0.694	0.816	0.944	0.983
	ROC AUC	0.825	0.924	0.958	0.949

Note. Classification performance metrics summarized across all time points with logistic model shown for reference.

Figure 64

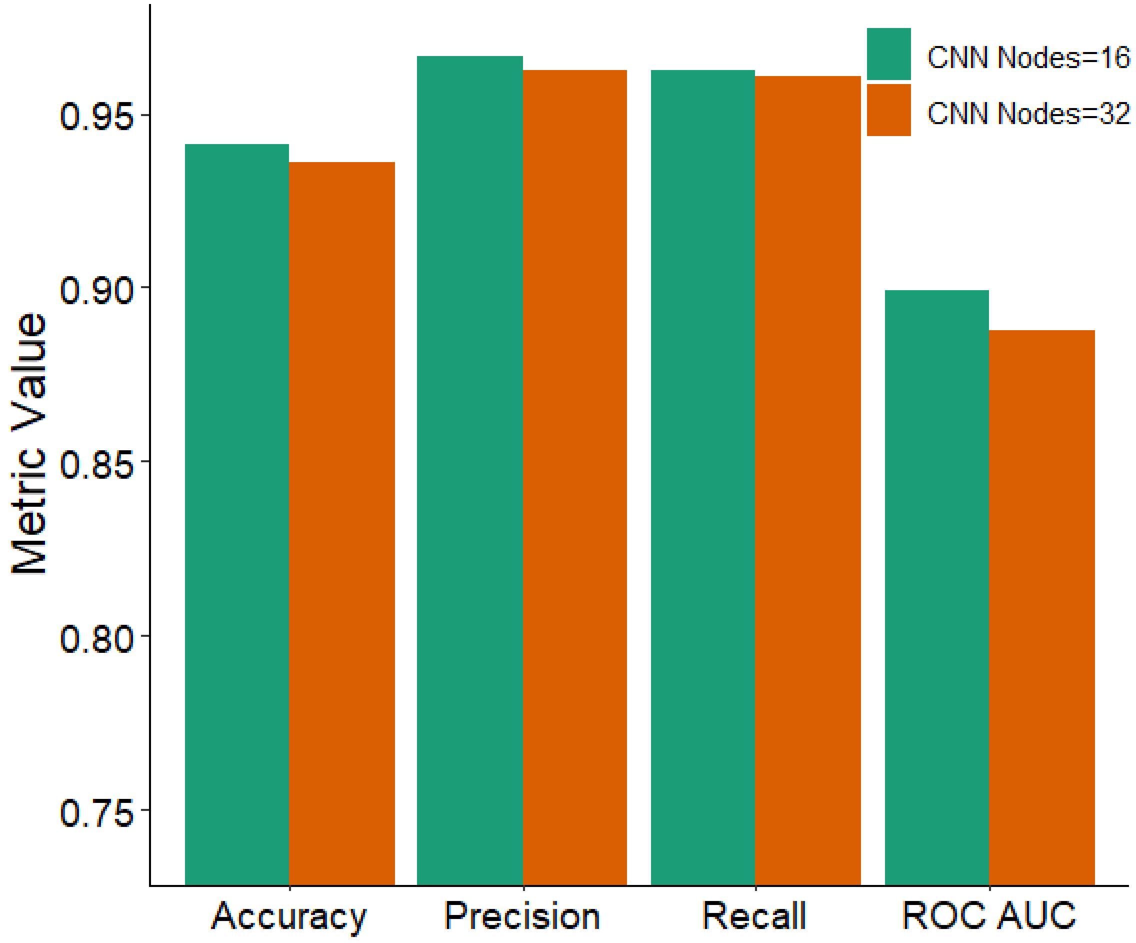
Feed-Forward Neural Network Hyperparameter Tunings – CDR-Based Impairment



Note. Hyperparameter tunings provided from cross-fold validations on whole subject trajectories.

Figure 65

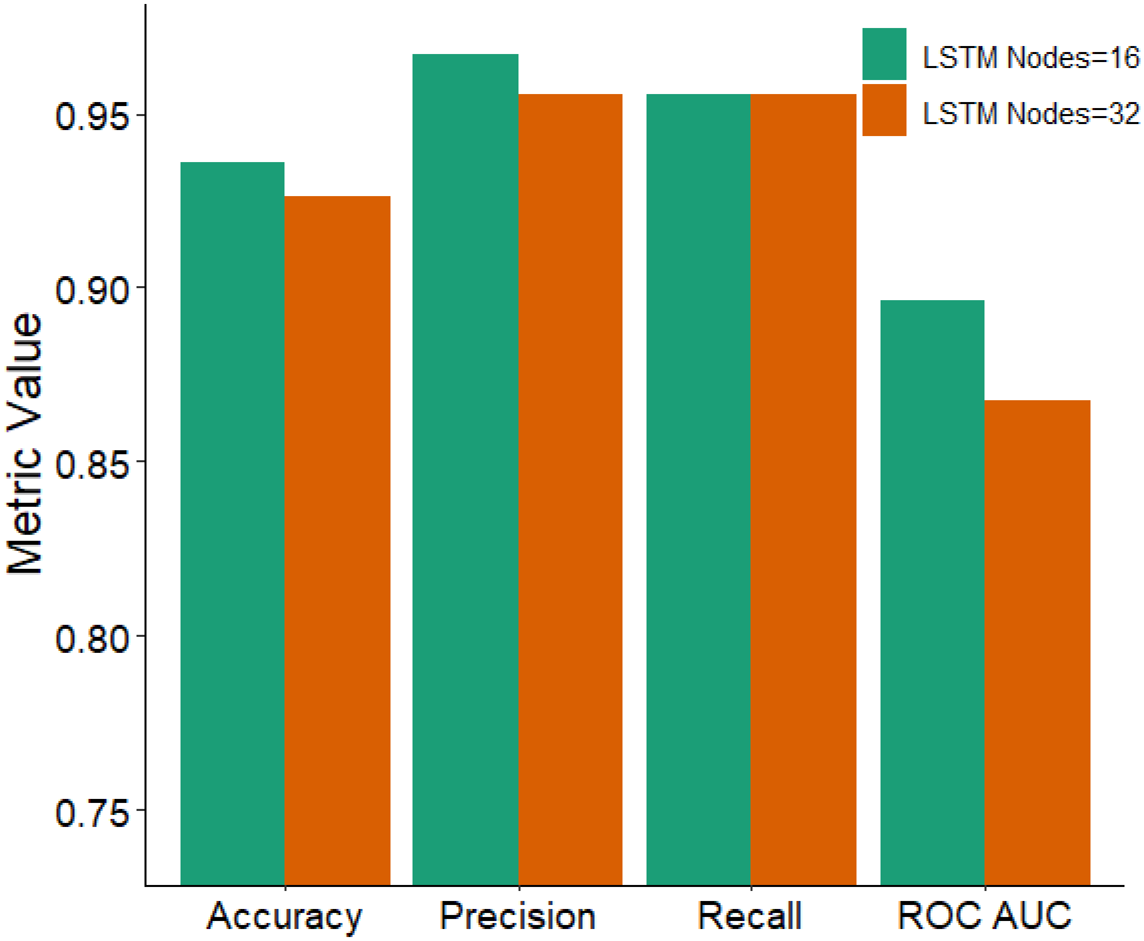
1D Convolutional Neural Network Hyperparameter Tunings – CDR-Based Impairment



Note. Hyperparameter tunings provided from cross-fold validations on whole subject trajectories.

Figure 66

LSTM Recurrent Neural Network Hyperparameter Tunings – CDR-Based Impairment



Note. Hyperparameter tunings provided from cross-fold validations on whole subject trajectories.

Table 58

Top Hyperparameters for Neural Networks – CDR-Based Impairment

Neural network	Performance metric	Best value	Hyperparameter set
Feed-forward neural network			
	Accuracy	0.821	3 layers: {8,4,2}
	Precision	0.935	2 layers: {4,2}
	Recall	0.844	3 layers: {8,4,2}
	ROC AUC	0.785	2 layers: {4,2}
1D convolutional neural network			
	Accuracy	0.941	CNN nodes=16
	Precision	0.967	CNN nodes=16
	Recall	0.962	CNN nodes=16
	ROC AUC	0.899	CNN nodes=16
LSTM recurrent neural network			
	Accuracy	0.936	LSTM nodes=16
	Precision	0.967	LSTM nodes=16
	Recall	0.956	LSTM nodes=32
	ROC AUC	0.896	LSTM nodes=16

Note. For each neural network, the best performing hyperparameter set for each classification metric is displayed along with the corresponding value.

Table 59

Feed-Forward Neural Network Prediction Performance – Whole Subject Trajectories of CDR-Based Impairment

Years	Accuracy	Precision	Recall	AUC	NRI
0.0	0.829	0.972	0.817	0.925	0.057
0.5	0.803	0.964	0.789	0.896	0.034
1.0	0.825	0.981	0.803	0.908	0.151
1.5	0.804	0.989	0.804	0.845	0.247
2.0	0.797	0.974	0.733	0.895	0.240
2.5	0.517	1.000	0.517	1.000	-
3.0	0.686	0.951	0.609	0.832	0.170
4.0	0.843	0.929	0.650	0.840	0.332
5.0	0.818	1.000	0.647	0.890	0.419
6.0	0.812	0.900	0.643	0.857	0.468
All	0.800	0.972	0.768	0.908	0.134

Note. Classification performance metrics presented at individual time points as well as across all time points. NRI based on reclassification from logistic reference model; NRI incalculable under certain ROC AUC conditions.

Table 60

Evaluation of Feed-Forward Neural Network Relative to Logistic Reference – Whole Subject Trajectories of CDR-Based Impairment

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
Accuracy	0.800	0.711	0.090	12.61%	[0.068, 0.111]
Precision	0.972	0.958	0.014	1.48%	[0.002, 0.025]
Recall	0.768	0.661	0.107	16.26%	[0.083, 0.133]
ROC AUC	0.908	0.841	0.066	7.89%	[0.051, 0.082]

Note. Values for change in metric and percent difference relative to logistic reference model; bootstrap confidence interval corresponds to change in metric.

Figure 67

Feed-Forward Neural Network True and Predicted CDR-Based Impairment Counts – Whole Subject Trajectories

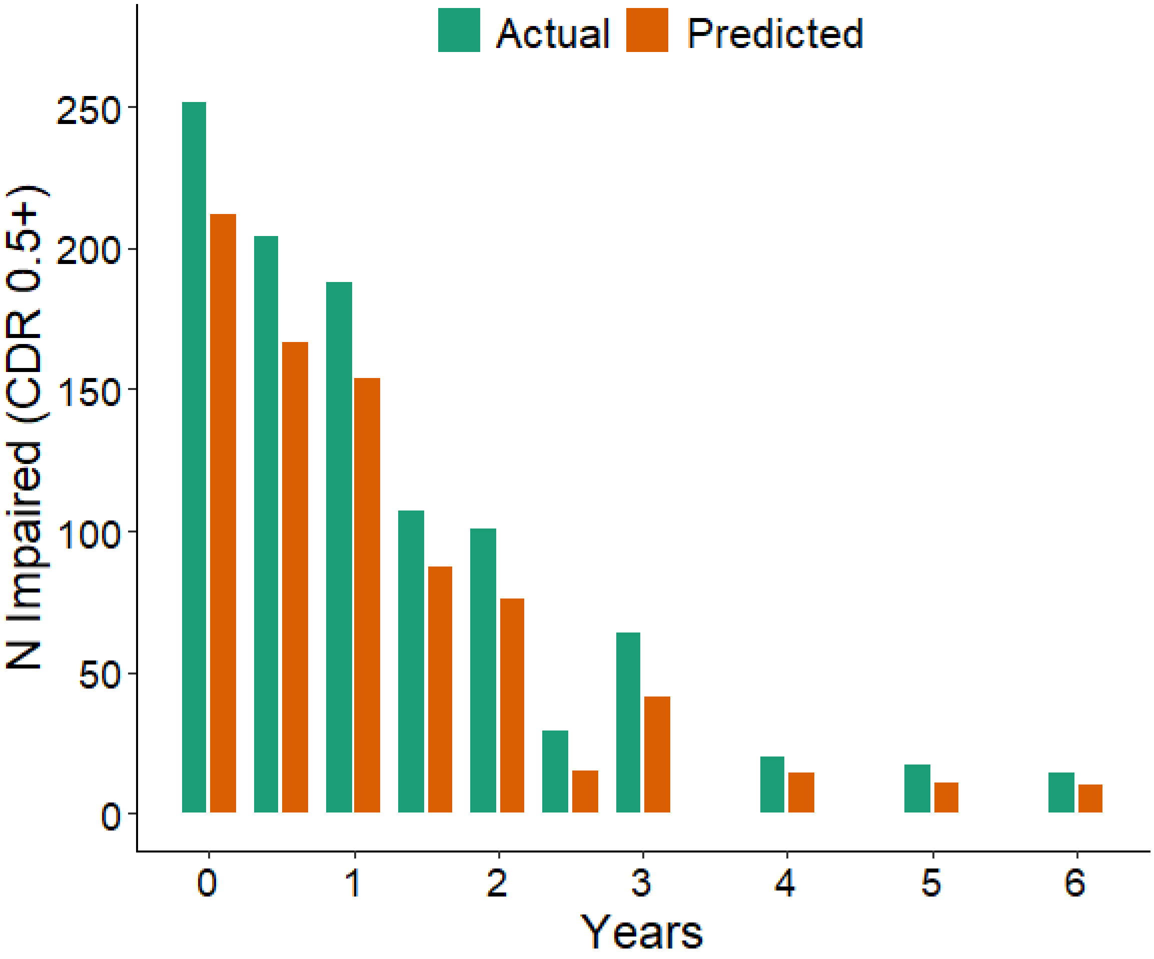
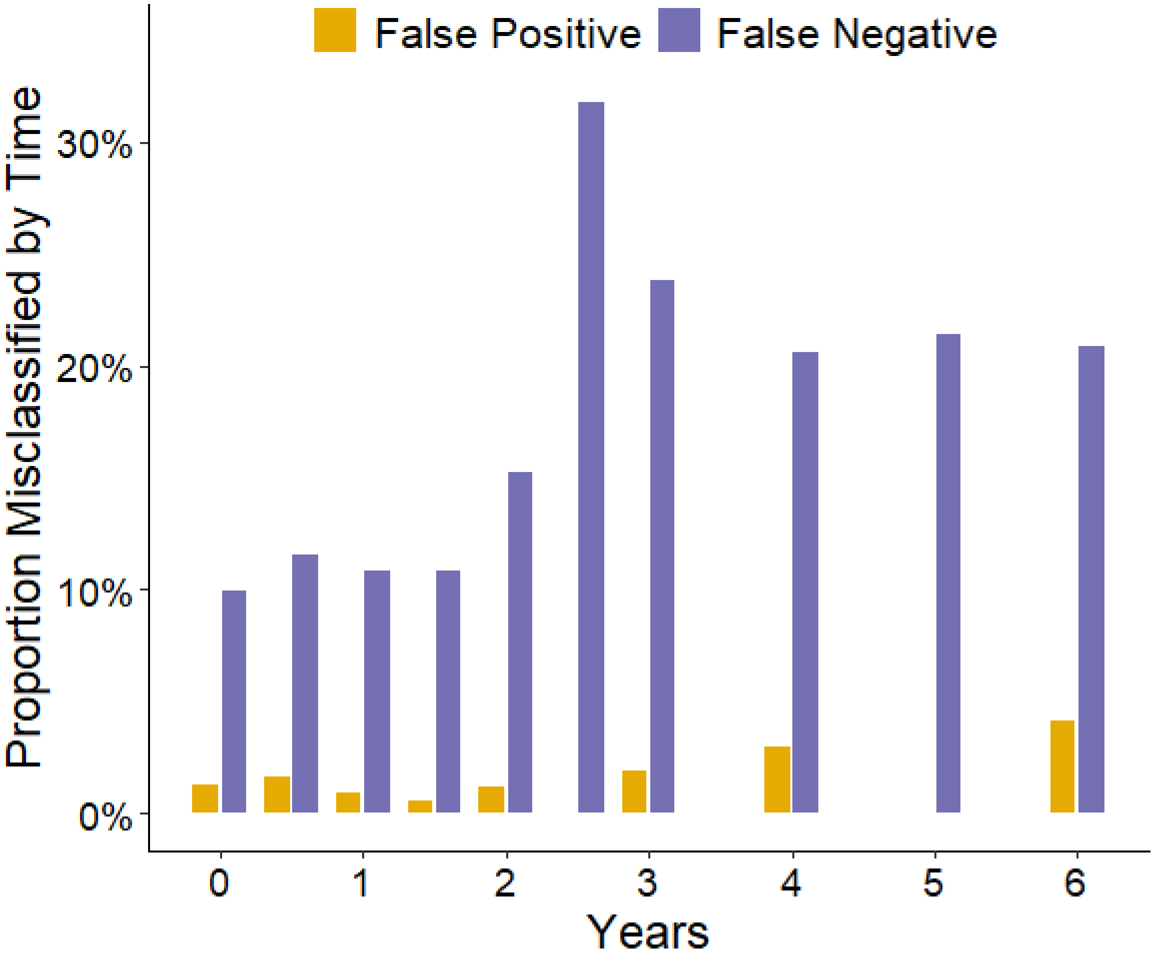


Figure 68

Feed-Forward Neural Network Misclassification Rates for CDR-Based Impairment – Whole Subject Trajectories



Note. False positive and false negative rates marginalized by time to be relative to total counts at that timepoint.

Table 61

Feed-Forward Neural Network Prediction Performance – Final Observation Forecasts of CDR-Based Impairment

Years	Accuracy	Precision	Recall	AUC	NRI
0.5	0.786	0.917	0.846	0.846	0.231
1.0	0.867	1.000	0.867	1.000	-
1.5	0.967	1.000	0.967	1.000	-
2.0	0.889	0.964	0.900	0.944	0.100
3.0	0.717	0.966	0.700	0.783	0.150
4.0	0.867	1.000	0.800	0.940	0.300
5.0	0.750	1.000	0.571	0.943	-0.143
6.0	0.806	1.000	0.684	0.908	0.263
All	0.832	0.980	0.816	0.924	0.123

Note. Classification performance metrics presented at individual time points as well as across all time points. NRI based on reclassification from logistic reference model; NRI incalculable under certain ROC AUC conditions.

Table 62

Evaluation of Feed-Forward Neural Network Relative to Logistic Reference – Final Observation Forecasts of CDR-Based Impairment

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
Accuracy	0.832	0.730	0.102	13.91%	[0.045, 0.148]
Precision	0.980	0.977	0.003	0.34%	[-0.022, 0.023]
Recall	0.816	0.694	0.121	17.45%	[0.060, 0.177]
ROC AUC	0.924	0.825	0.099	12.00%	[0.058, 0.131]

Note. Values for change in metric and percent difference relative to logistic reference model; bootstrap confidence interval corresponds to change in metric.

Figure 69

Feed-Forward Neural Network True and Predicted CDR-Based Impairment Counts – Final Observation Forecasts

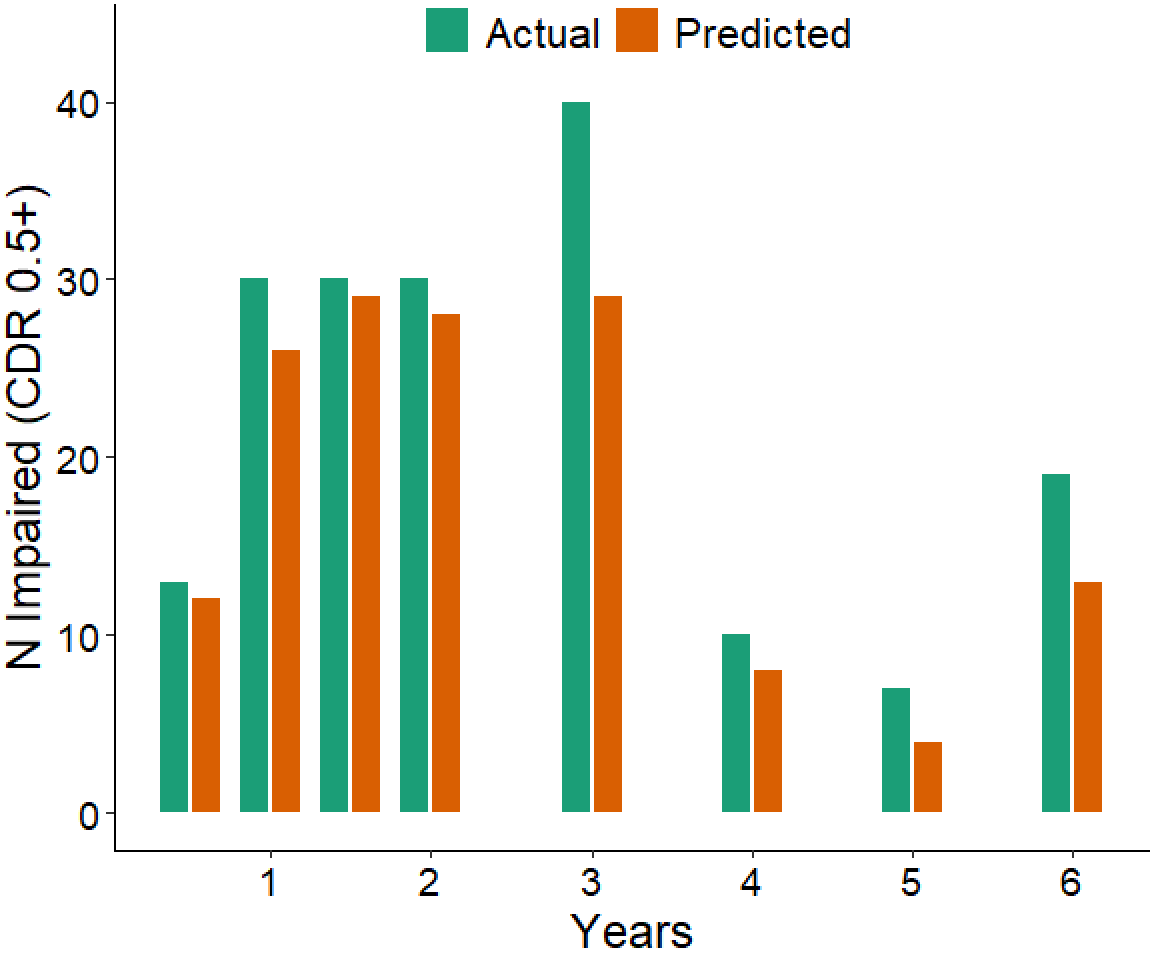
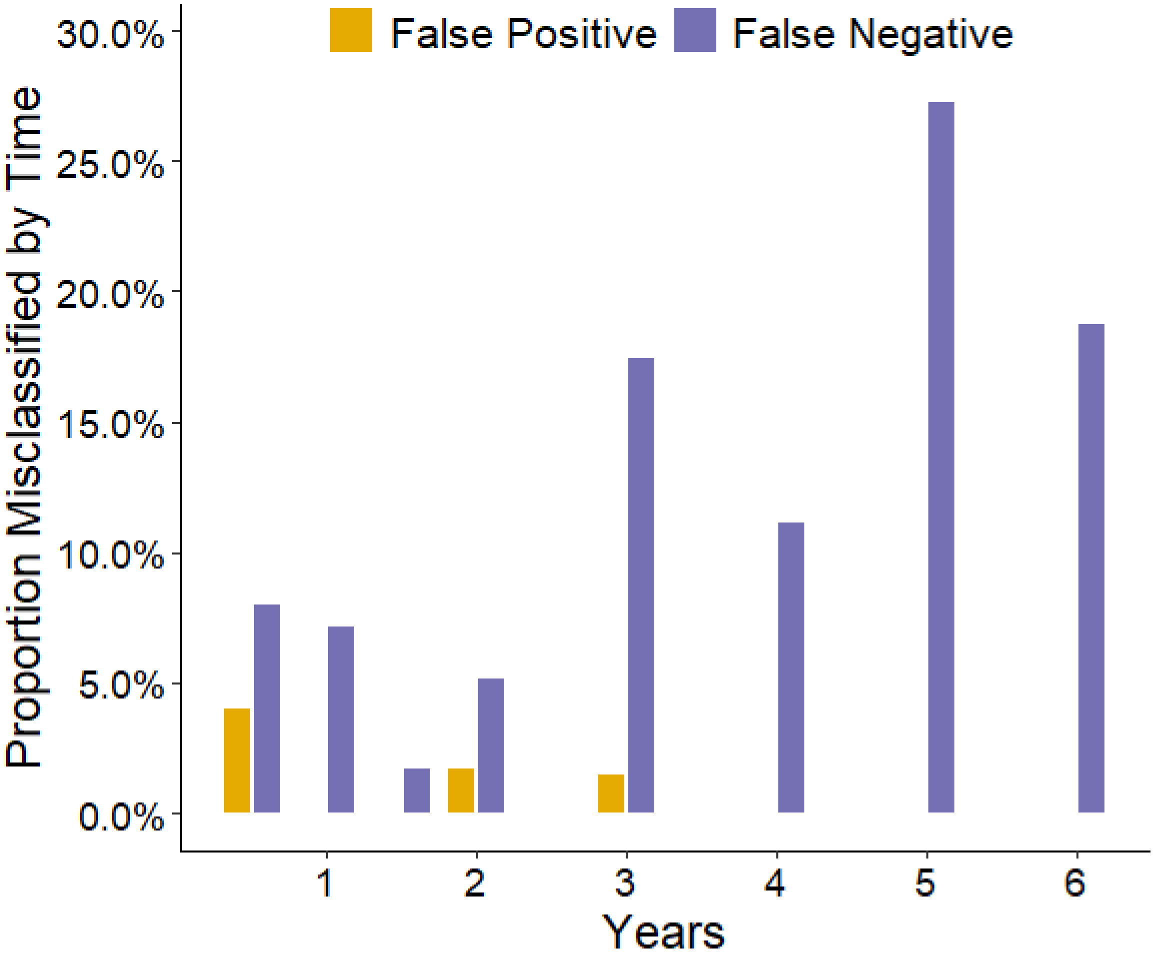


Figure 70

Feed-Forward Neural Network Misclassification Rates for CDR-Based Impairment – Final Observation Forecasts



Note. False positive and false negative rates marginalized by time to be relative to total counts at that timepoint.

Table 63

1D Convolutional Neural Network Prediction Performance – Whole Subject Trajectories of CDR-Based Impairment

Years	Accuracy	Precision	Recall	AUC	NRI
0.0	0.870	0.973	0.864	0.879	0.113
0.5	0.851	0.966	0.848	0.857	0.093
1.0	0.855	0.970	0.851	0.863	0.149
1.5	0.857	0.960	0.888	0.544	-0.269
2.0	0.797	0.939	0.762	0.822	0.198
2.5	0.724	1.000	0.724	1.000	-
3.0	0.744	0.957	0.688	0.798	0.249
4.0	0.843	0.929	0.650	0.809	0.332
5.0	0.818	0.923	0.706	0.822	0.415
6.0	0.781	0.889	0.571	0.758	0.397
All	0.836	0.963	0.822	0.855	0.159

Note. Classification performance metrics presented at individual time points as well as across all time points. NRI based on reclassification from logistic reference model; NRI incalculable under certain ROC AUC conditions.

Table 64

Evaluation of 1D Convolutional Neural Network Relative to Logistic Reference – Whole Subject Trajectories of CDR-Based Impairment

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
Accuracy	0.836	0.711	0.126	17.68%	[0.105, 0.147]
Precision	0.963	0.958	0.005	0.53%	[-0.009, 0.018]
Recall	0.822	0.661	0.161	24.44%	[0.137, 0.186]
ROC AUC	0.855	0.841	0.013	1.60%	[0.022, 0.066]

Note. Values for change in metric and percent difference relative to logistic reference model; bootstrap confidence interval corresponds to change in metric.

Figure 71

1D Convolutional Neural Network True and Predicted CDR-Based Impairment Counts – Whole Subject Trajectories

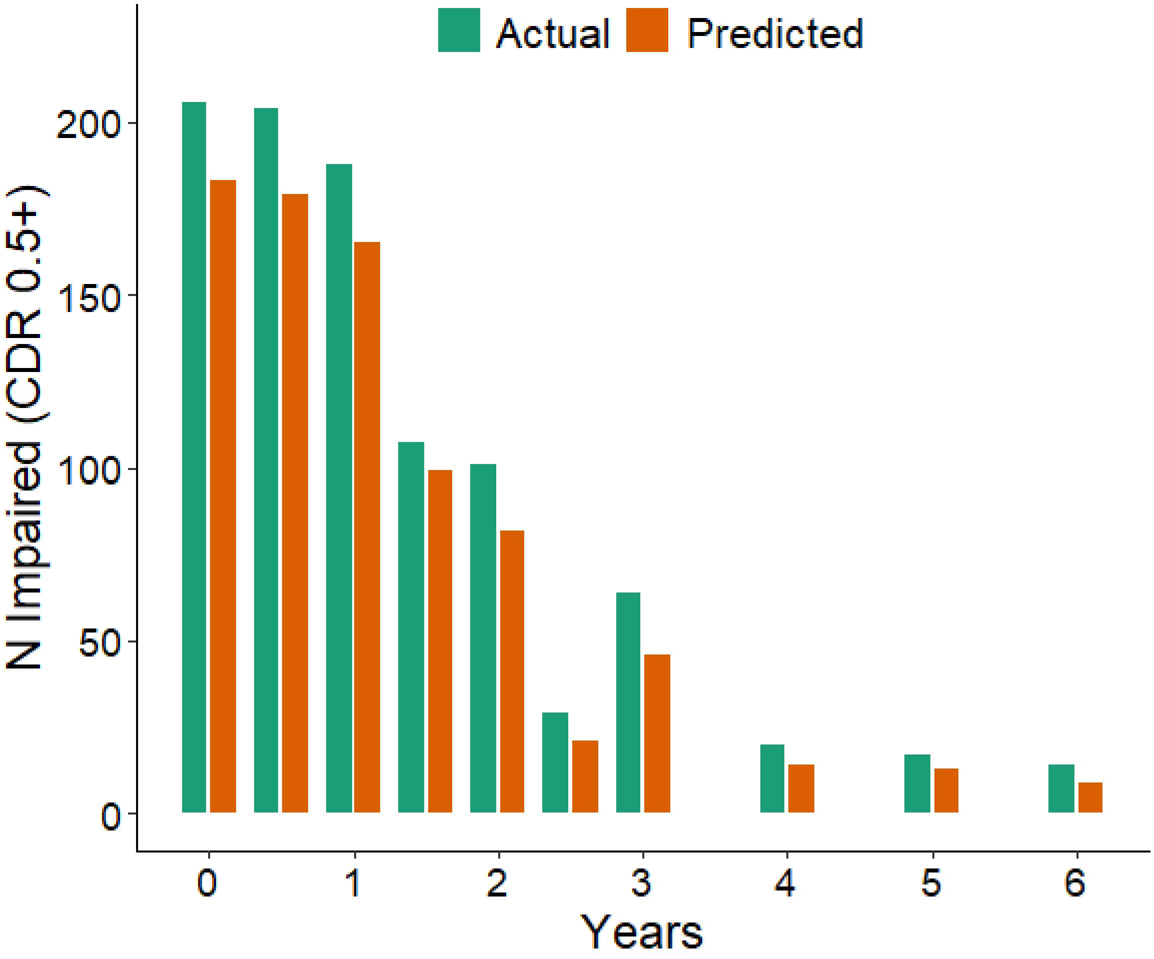
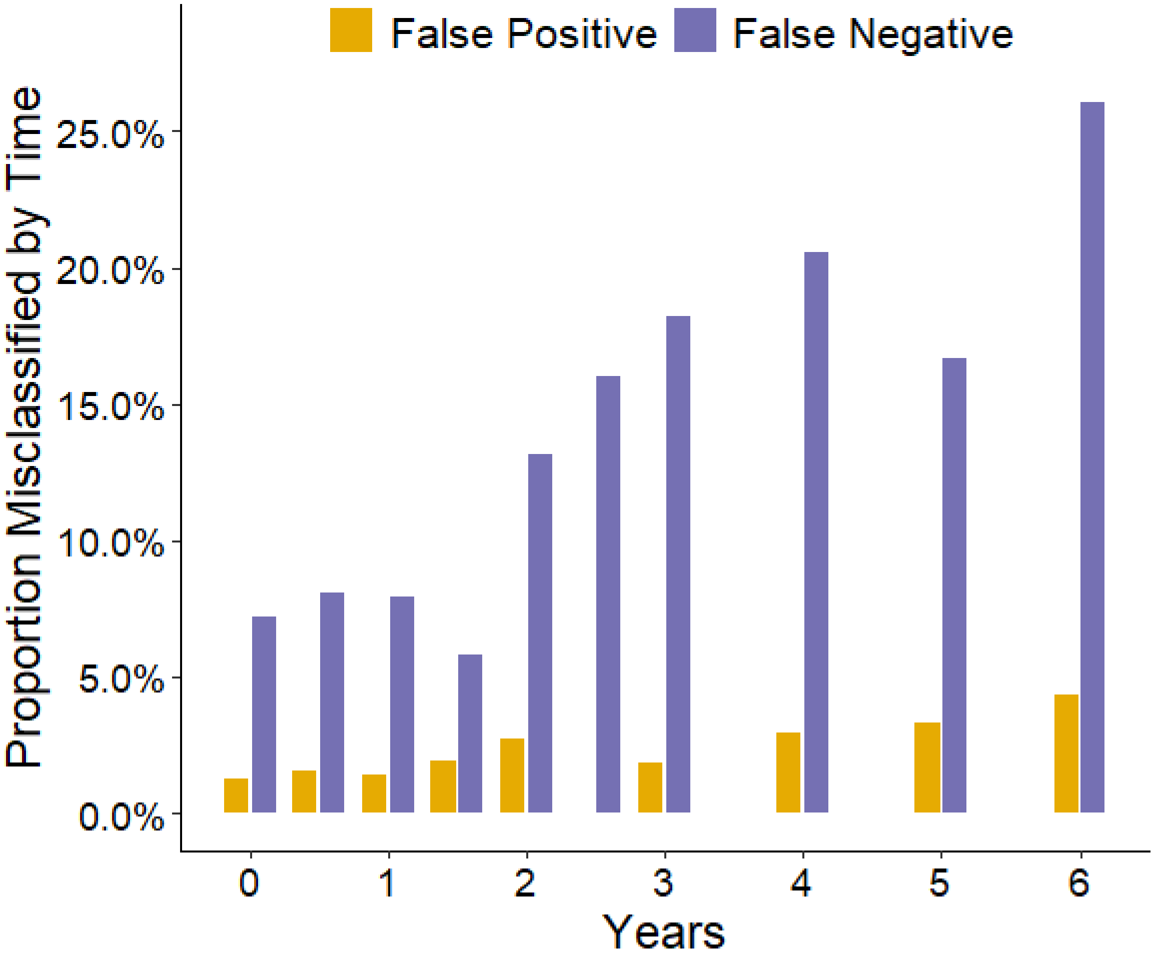


Figure 72

1D Convolutional Neural Network Misclassification Rates for CDR-Based Impairment – Whole Subject Trajectories



Note. False positive and false negative rates marginalized by time to be relative to total counts at that timepoint.

Table 65

1D Convolutional Neural Network Prediction Performance – Final Observation Forecasts of CDR-Based Impairment

Years	Accuracy	Precision	Recall	AUC	NRI
0.5	0.929	1.000	0.923	0.962	0.231
1.0	1.000	1.000	1.000	1.000	-
1.5	0.967	1.000	0.967	1.000	-
2.0	0.972	1.000	0.967	0.983	0.867
3.0	0.957	0.975	0.975	0.904	0.692
4.0	1.000	1.000	1.000	1.000	-
5.0	0.750	1.000	0.571	0.786	0.600
6.0	0.903	1.000	0.842	0.921	0.575
All	0.949	0.994	0.944	0.958	0.787

Note. Classification performance metrics presented at individual time points as well as across all time points. NRI based on reclassification from logistic reference model; NRI incalculable under certain ROC AUC conditions.

Table 66

Evaluation of 1D Convolutional Neural Network Relative to Logistic Reference – Final Observation Forecasts of CDR-Based Impairment

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
Accuracy	0.949	0.730	0.218	29.90%	[0.194, 0.248]
Precision	0.994	0.977	0.018	1.80%	[0.005, 0.023]
Recall	0.944	0.694	0.250	35.96%	[0.222, 0.284]
ROC AUC	0.958	0.825	0.133	16.15%	[0.147, 0.174]

Note. Values for change in metric and percent difference relative to logistic reference model; bootstrap confidence interval corresponds to change in metric.

Figure 73

1D Convolutional Neural Network True and Predicted CDR-Based Impairment Counts – Final Observation Forecasts

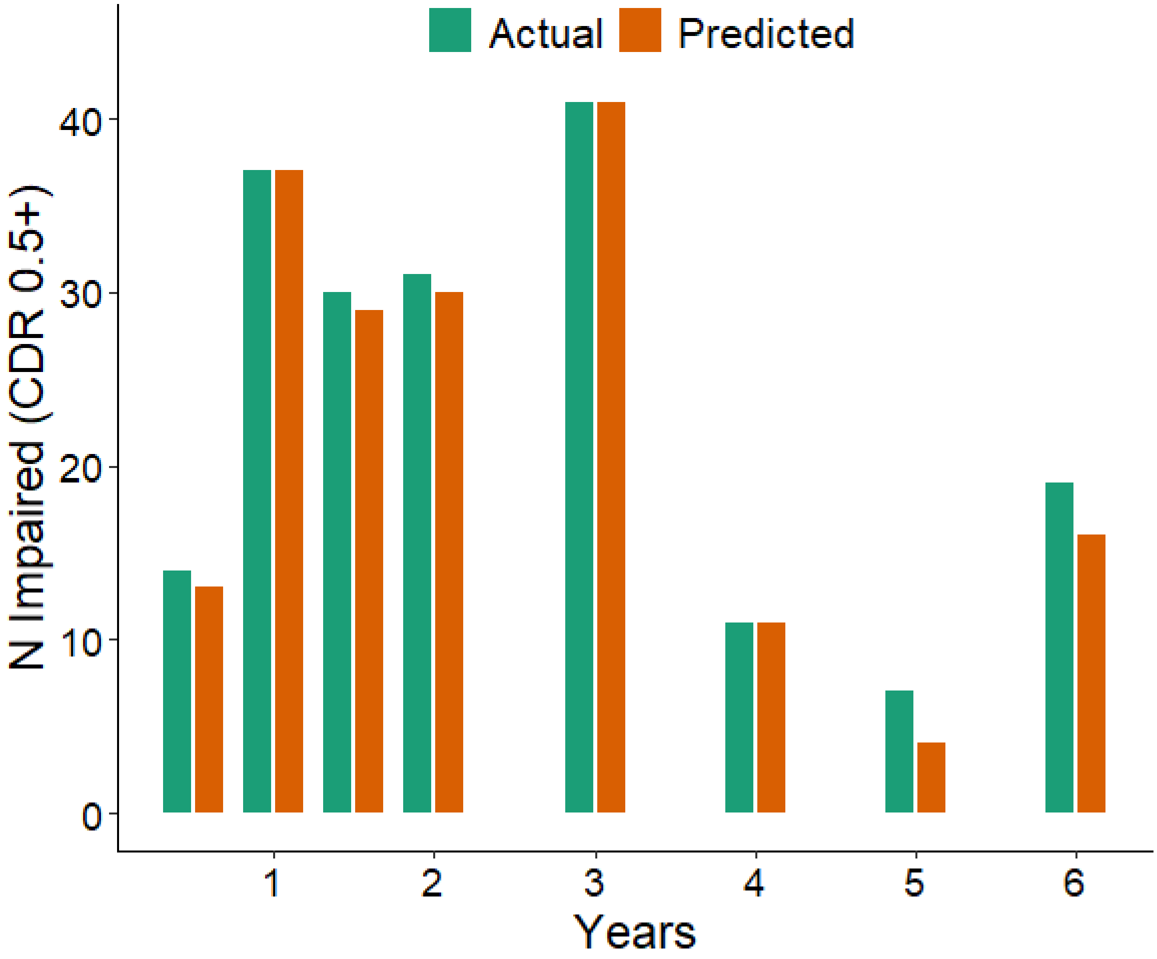
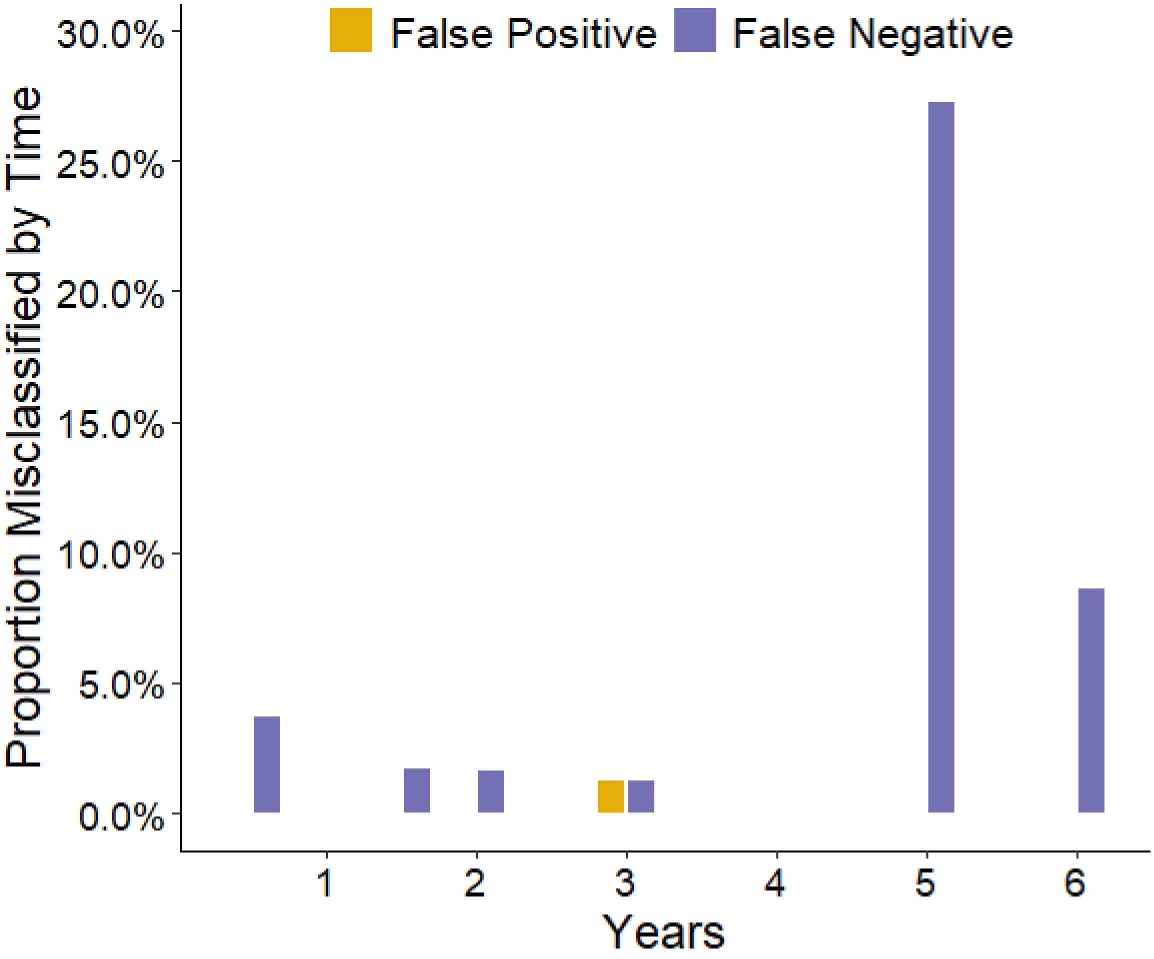


Figure 74

1D Convolutional Neural Network Misclassification Rates for CDR-Based Impairment – Final Observation Forecasts



Note. False positive and false negative rates marginalized by time to be relative to total counts at that timepoint.

Table 67

LSTM Recurrent Neural Network Prediction Performance – Whole Subject Trajectories of CDR-Based Impairment

Years	Accuracy	Precision	Recall	AUC	NRI
0.0	0.877	0.978	0.869	0.892	0.139
0.5	0.855	0.967	0.853	0.860	0.098
1.0	0.855	0.964	0.856	0.853	0.129
1.5	0.857	0.960	0.888	0.544	-0.269
2.0	0.776	0.937	0.733	0.807	0.168
2.5	0.655	1.000	0.655	1.000	-
3.0	0.744	0.938	0.703	0.783	0.219
4.0	0.843	0.875	0.700	0.818	0.350
5.0	0.818	0.867	0.765	0.820	0.412
6.0	0.781	0.818	0.643	0.766	0.413
All	0.835	0.958	0.824	0.848	0.146

Note. Classification performance metrics presented at individual time points as well as across all time points. NRI based on reclassification from logistic reference model; NRI incalculable under certain ROC AUC conditions.

Table 68

Evaluation of LSTM Recurrent Neural Network Relative to Logistic Reference – Whole Subject Trajectories of CDR-Based Impairment

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
Accuracy	0.835	0.711	0.124	17.45%	[0.103, 0.144]
Precision	0.958	0.958	0.000	0.05%	[-0.013, 0.013]
Recall	0.824	0.661	0.164	24.76%	[0.139, 0.188]
ROC AUC	0.848	0.841	0.007	0.84%	[0.031, 0.072]

Note. Values for change in metric and percent difference relative to logistic reference model; bootstrap confidence interval corresponds to change in metric.

Figure 75

LSTM Recurrent Neural Network True and Predicted CDR-Based Impairment Counts – Whole Subject Trajectories

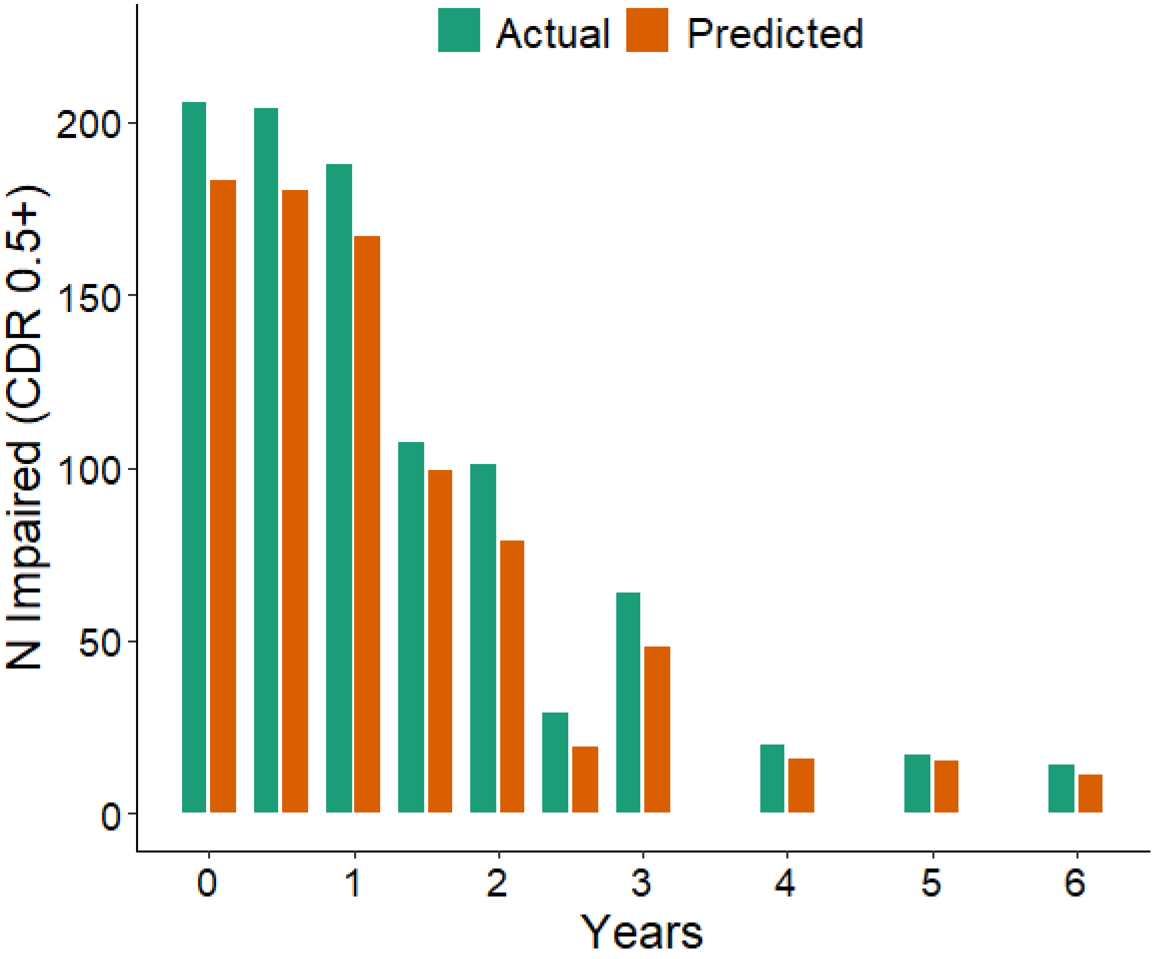
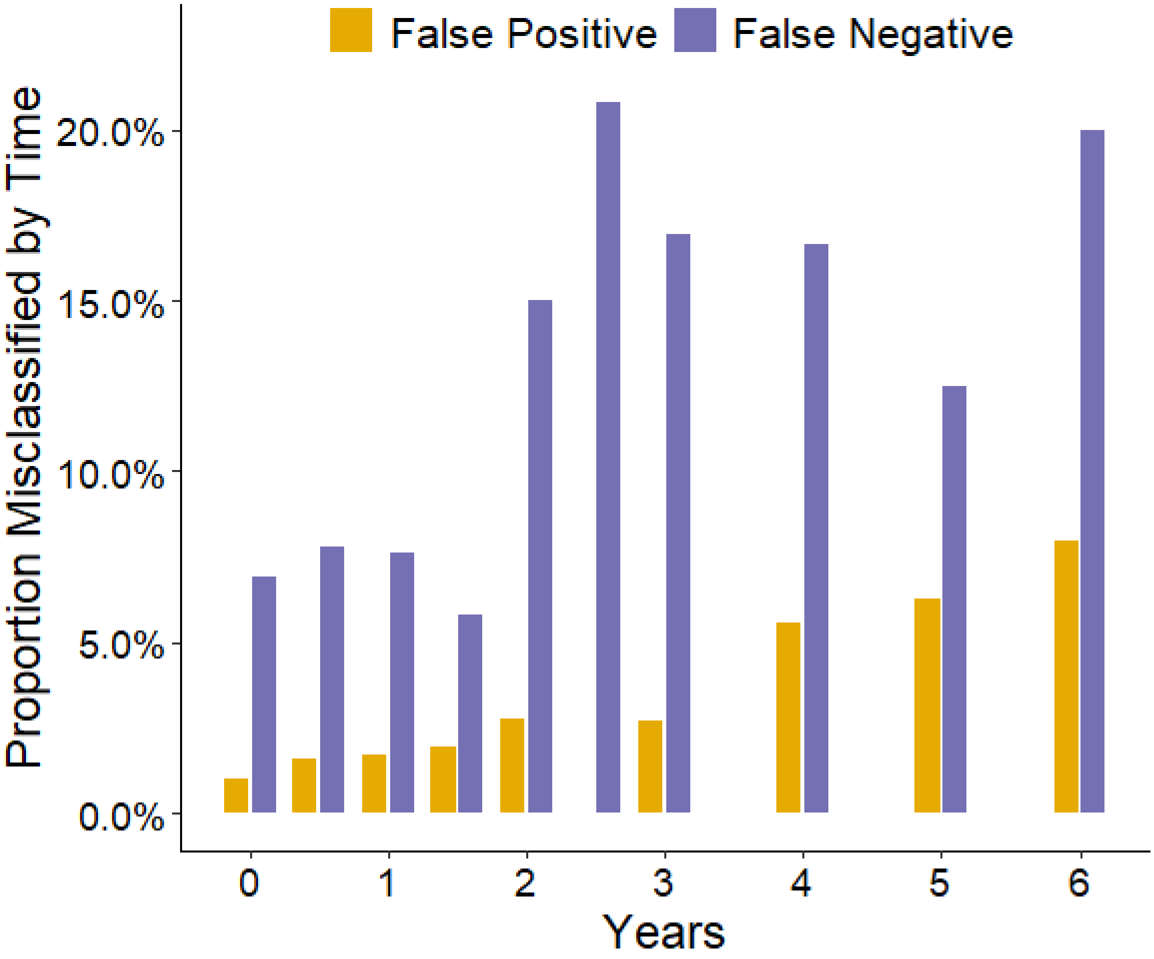


Figure 76

LSTM Recurrent Neural Network Misclassification Rates for CDR-Based Impairment – Whole Subject Trajectories



Note. False positive and false negative rates marginalized by time to be relative to total counts at that timepoint.

Table 69

LSTM Recurrent Neural Network Prediction Performance – Final Observation Forecasts of CDR-Based Impairment

Years	Accuracy	Precision	Recall	AUC	NRI
0.5	0.929	1.000	0.923	0.962	0.231
1.0	1.000	1.000	1.000	1.000	-
1.5	1.000	1.000	1.000	1.000	-
2.0	1.000	1.000	1.000	1.000	-
3.0	0.978	0.976	1.000	0.917	0.717
4.0	0.933	0.909	1.000	0.900	0.800
5.0	0.917	1.000	0.857	0.929	0.886
6.0	0.935	0.947	0.947	0.932	0.596
All	0.972	0.983	0.983	0.949	0.769

Note. Classification performance metrics presented at individual time points as well as across all time points. NRI based on reclassification from logistic reference model; NRI incalculable under certain ROC AUC conditions.

Table 70

Evaluation of LSTM Recurrent Neural Network Relative to Logistic Reference – Final Observation Forecasts of CDR-Based Impairment

Performance metric	Learning model	Reference model	Change in metric	Percent difference	Bootstrap 95% CI
Accuracy	0.972	0.730	0.242	33.10%	[0.221, 0.261]
Precision	0.983	0.977	0.007	0.68%	[-0.012, 0.023]
Recall	0.983	0.694	0.289	41.59%	[0.268, 0.306]
ROC AUC	0.949	0.825	0.124	15.06%	[0.133, 0.172]

Note. Values for change in metric and percent difference relative to logistic reference model; bootstrap confidence interval corresponds to change in metric.

Figure 77

LSTM Recurrent Neural Network True and Predicted CDR-Based Impairment Counts – Final Observation Forecasts

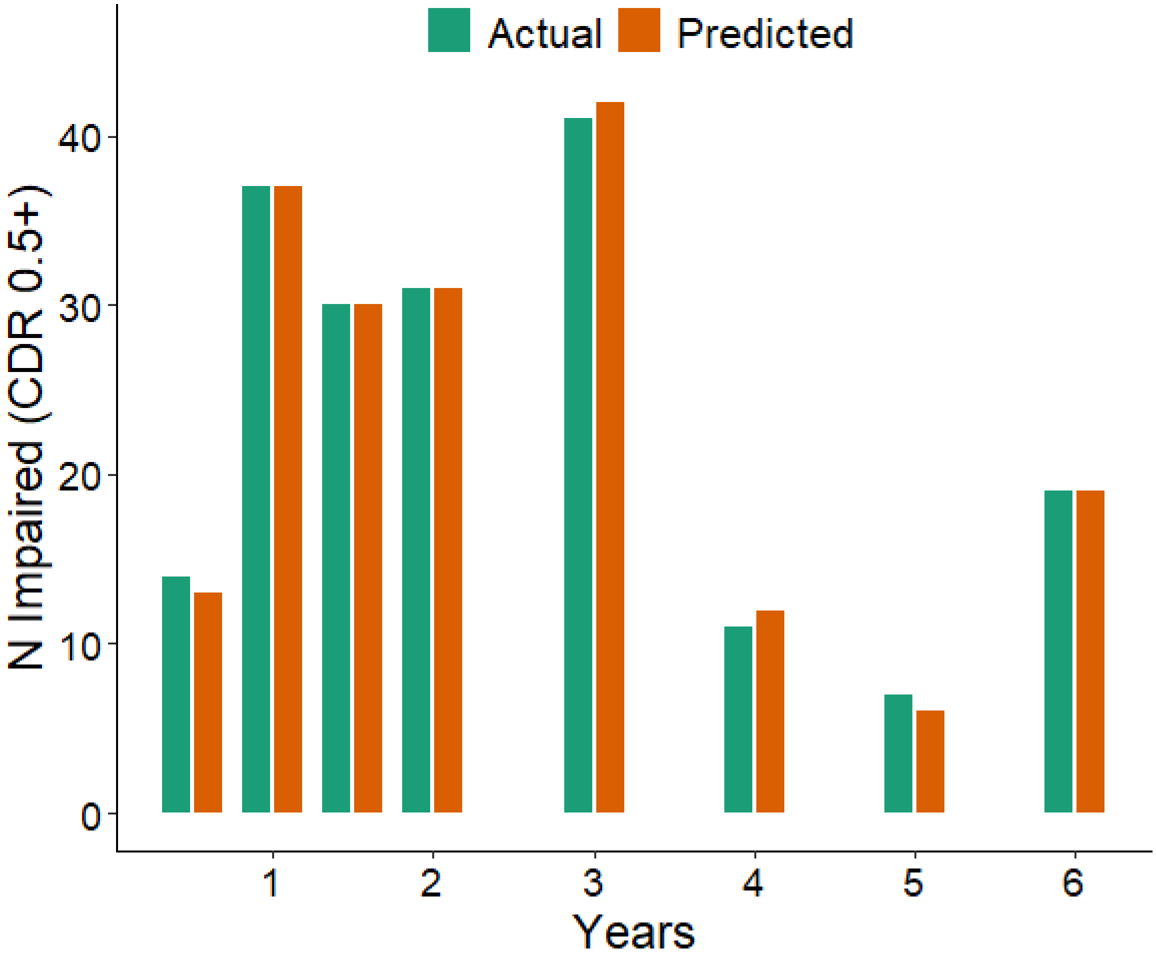
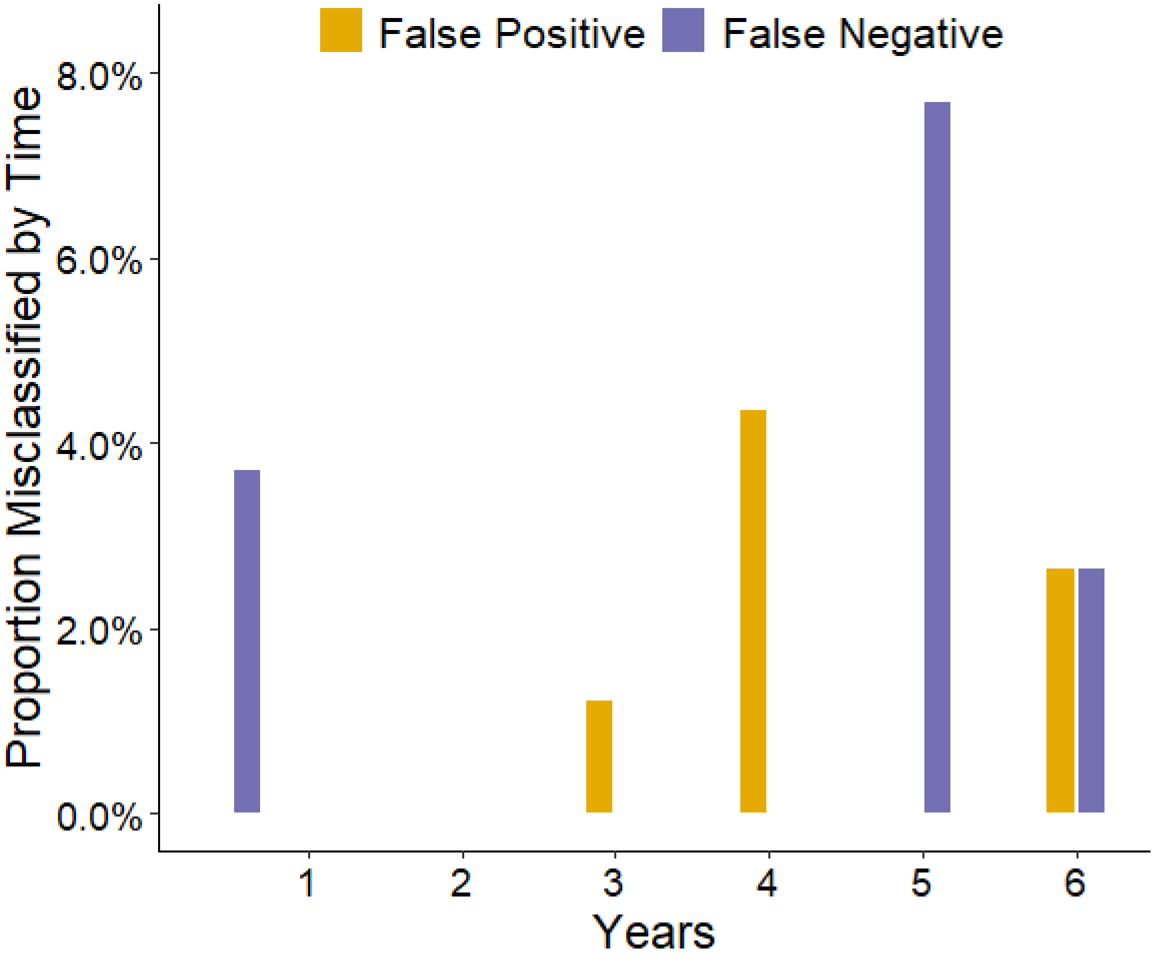


Figure 78

LSTM Recurrent Neural Network Misclassification Rates for CDR-Based Impairment – Final Observation Forecasts



Note. False positive and false negative rates marginalized by time to be relative to total counts at that timepoint.

Table 71

Cross-Model Comparisons – Root Mean Square Error of ADAS-Cog Score for Whole Subject Trajectories

Model	Reference	FNN	MERF	Boosted	Bagged	GLMM	LSTM	CNN
Reference	6.819	-2.160 (-31.67%) <.001	-1.891 (-27.73%) <.001	-1.876 (-27.51%) <.001	-1.277 (-18.73%) <.001	-1.135 (-16.64%) <.001	-0.652 (-9.56%) <.001	-0.313 (-4.59%) .041
FNN	[-2.351, -1.972]	4.660	-0.269 (-5.45%) .004	-0.284 (-5.74%) .007	-0.882 (-15.92%) <.001	-1.025 (-18.03%) <.001	-1.508 (-24.45%) <.001	-1.847 (-28.38%) <.001
MERF	[-2.112, -1.658]	[-0.469, -0.062]	4.928	-0.015 (-0.31%) .457	-0.614 (-11.07%) <.001	-0.756 (-13.30%) <.001	-1.239 (-20.09%) <.001	-1.578 (-24.26%) <.001
Boosted	[-2.097, -1.662]	[-0.481, -0.069]	[-0.247, 0.207]	4.943	-0.598 (-10.80%) <.001	-0.741 (-13.03%) <.001	-1.224 (-19.84%) <.001	-1.563 (-24.02%) <.001
Bagged	[-1.503, -1.041]	[-1.102, -0.675]	[-0.824, -0.393]	[-0.813, -0.387]	5.542	-0.142 (-2.51%) .107	-0.625 (-10.14%) <.001	-0.965 (-14.83%) <.001
GLMM	[-1.378, -0.890]	[-1.212, -0.831]	[-0.977, -0.523]	[-0.951, -0.531]	[-0.387, 0.093]	5.684	-0.483 (-7.83%) <.001	-0.822 (-12.64%) <.001
LSTM	[-0.996, -0.260]	[-1.699, -1.298]	[-1.463, -1.018]	[-1.429, -1.031]	[-0.846, -0.395]	[-0.718, -0.245]	6.167	-0.339 (-5.21%) .034
CNN	[-0.705, 0.050]	[-2.054, -1.655]	[-1.800, -1.341]	[-1.773, -1.370]	[-1.181, -0.740]	[-1.053, -0.572]	[-0.682, 0.013]	6.506

Note. Whole subject trajectory RMSE values across all timepoints for each model listed on the diagonal, beginning with the CPath reference, followed by the top performing learning model. Learning models are then ordered according to decreasing performance. Upper half of matrix contains numeric difference in metric, with percent difference in parenthesis, then one-sided bootstrapped proportional *p*-value; lower half of matrix contains the corresponding bootstrapped 95% confidence interval for the numeric difference. All differences are for the better performing model in the column relative to the poorer performing model in the row.

Table 72

Cross-Model Comparisons – Root Mean Square Error of ADAS-Cog Score for Final Observation Forecasts

Model	Reference	GLMM	MERF	Bagged	Boosted	FNN	LSTM	CNN
Reference	9.720	-5.192 (-53.42%) <.001	-4.999 (-51.43%) <.001	-4.577 (-47.09%) <.001	-4.487 (-46.16%) <.001	-4.237 (-43.59%) <.001	-4.070 (-41.88%) <.001	-3.864 (-39.76%) <.001
GLMM	[-5.767, -4.665]	4.528	-0.193 (-4.09%) .235	-0.615 (-11.95%) .019	-0.705 (-13.47%) .010	-0.955 (-17.42%) .001	-1.122 (-19.86%) <.001	-1.328 (-22.68%) <.001
MERF	[-5.490, -4.532]	[-0.754, 0.377]	4.721	-0.421 (-8.20%) .049	-0.512 (-9.78%) .021	-0.762 (-13.90%) .001	-0.929 (-16.44%) .001	-1.135 (-19.38%) <.001
Bagged	[-5.154, -4.026]	[-1.170, -0.030]	[-0.934, 0.095]	5.142	-0.090 (-1.73%) .383	-0.340 (-6.21%) .120	-0.507 (-8.98%) .046	-0.713 (-12.18%) .011
Boosted	[-5.073, -3.937]	[-1.289, -0.101]	[-1.004, -0.006]	[-0.677, 0.544]	5.233	-0.250 (-4.56%) .190	-0.417 (-7.38%) .076	-0.623 (-10.63%) .024
FNN	[-4.814, -3.699]	[-1.558, -0.380]	[-1.284, -0.265]	[-0.950, 0.208]	[-0.858, 0.303]	5.483	-0.167 (-2.95%) .270	-0.373 (-6.37%) .116
LSTM	[-5.099, -3.156]	[-1.682, -0.539]	[-1.418, -0.408]	[-1.131, 0.072]	[-1.019, 0.151]	[-0.748, 0.393]	5.650	-0.206 (-3.52%) .304
CNN	[-4.822, -2.844]	[-1.886, -0.728]	[-1.693, -0.611]	[-1.333, -0.118]	[-1.197, -0.014]	[-0.906, 0.218]	[-1.114, 0.699]	5.856

Note. Final observation forecast RMSE values across all timepoints for each model listed on the diagonal, beginning with the CPath reference, followed by the top performing learning model. Learning models are then ordered according to decreasing performance. Upper half of matrix contains numeric difference in metric, with percent difference in parenthesis, then one-sided bootstrapped proportional *p*-value; lower half of matrix contains the corresponding bootstrapped 95% confidence interval for the numeric difference. All differences are for the better performing model in the column relative to the poorer performing model in the row.

Table 73

Cross-Model Comparisons – Mean Absolute Error of ADAS-Cog Score for Whole Subject Trajectories

Model	Reference	FNN	MERF	Boosted	Bagged	LSTM	GLMM	CNN
Reference	5.283	-1.723 (-32.61%) <.001	-1.491 (-28.23%) <.001	-1.426 (-26.98%) <.001	-0.960 (-18.18%) <.001	-0.861 (-16.29%) <.001	-0.856 (-16.20%) <.001	-0.610 (-11.54%) <.001
FNN	[-1.876, -1.579]	3.560	-0.232 (-6.11%) .004	-0.297 (-7.71%) <.001	-0.763 (-17.64%) <.001	-0.862 (-19.50%) <.001	-0.867 (-19.58%) <.001	-1.113 (-23.82%) <.001
MERF	[-1.651, -1.321]	[-0.374, -0.078]	3.792	-0.066 (-1.71%) .211	-0.531 (-12.29%) <.001	-0.631 (-14.26%) <.001	-0.635 (-14.35%) <.001	-0.882 (-18.86%) <.001
Boosted	[-1.590, -1.275]	[-0.444, -0.143]	[-0.232, 0.087]	3.858	-0.465 (-10.76%) <.001	-0.565 (-12.77%) <.001	-0.570 (-12.87%) <.001	-0.816 (-17.45%) <.001
Bagged	[-1.131, -0.786]	[-0.921, -0.603]	[-0.682, -0.371]	[-0.626, -0.308]	4.323	-0.100 (-2.25%) .124	-0.104 (-2.35%) .112	-0.350 (-7.50%) <.001
LSTM	[-1.067, -0.629]	[-1.007, -0.697]	[-0.775, -0.478]	[-0.717, -0.418]	[-0.268, 0.084]	4.423	-0.005 (-0.10%) .490	-0.251 (-5.37%) .010
GLMM	[-1.031, -0.681]	[-1.015, -0.716]	[-0.795, -0.465]	[-0.731, -0.407]	[-0.273, 0.073]	[-0.217, 0.232]	4.427	-0.246 (-5.27%) .002
CNN	[-0.845, -0.397]	[-1.267, -0.963]	[-1.032, -0.715]	[-0.961, -0.658]	[-0.521, -0.183]	[-0.464, -0.036]	[-0.422, -0.061]	4.673

Note. Whole subject trajectory MAE values across all timepoints for each model listed on the diagonal, beginning with the CPath reference, followed by the top performing learning model. Learning models are then ordered according to decreasing performance. Upper half of matrix contains numeric difference in metric, with percent difference in parenthesis, then one-sided bootstrapped proportional *p*-value; lower half of matrix contains the corresponding bootstrapped 95% confidence interval for the numeric difference. All differences are for the better performing model in the column relative to the poorer performing model in the row.

Table 74

Cross-Model Comparisons – Mean Absolute Error of ADAS-Cog Score for Final Observation Forecasts

Model	Reference	GLMM	MERF	LSTM	CNN	Bagged	Boosted	FNN
Reference	6.859	-3.501 (-51.05%) <.001	-3.271 (-47.69%) <.001	-3.132 (-45.67%) <.001	-3.035 (-44.26%) <.001	-3.017 (-43.98%) <.001	-2.888 (-42.10%) <.001	-2.602 (-37.94%) <.001
GLMM	[-3.885, -3.105]	3.357	-0.231 (-6.43%) .126	-0.369 (-9.91%) .039	-0.466 (-12.19%) .016	-0.485 (-12.61%) .004	-0.614 (-15.45%) .004	-0.899 (-21.13%) <.001
MERF	[-3.635, -2.899]	[-0.616, 0.175]	3.588	-0.139 (-3.72%) .247	-0.235 (-6.15%) .125	-0.254 (-6.61%) .097	-0.383 (-9.64%) .029	-0.669 (-15.71%) .002
LSTM	[-3.681, -2.578]	[-0.735, 0.043]	[-0.512, 0.273]	3.727	-0.097 (-2.53%) .352	-0.115 (-3.00%) .321	-0.244 (-6.15%) .177	-0.530 (-12.46%) .035
CNN	[-3.528, -2.449]	[-0.856, -0.043]	[-0.653, 0.168]	[-0.605, 0.441]	3.823	-0.019 (-0.49%) .462	-0.148 (-3.72%) .260	-0.434 (-10.18%) .074
Bagged	[-3.464, -2.601]	[-0.864, -0.096]	[-0.658, 0.119]	[-0.630, 0.467]	[-0.577, 0.596]	3.842	-0.129 (-3.25%) .283	-0.415 (-9.75%) .031
Boosted	[-3.317, -2.463]	[-0.997, -0.212]	[-0.763, 0.014]	[-0.748, 0.271]	[-0.698, 0.416]	[-0.548, 0.354]	3.971	-0.286 (-6.72%) .093
FNN	[-3.051, -2.182]	[-1.280, -0.509]	[-1.071, -0.293]	[-1.091, 0.041]	[-0.984, 0.194]	[-0.831, 0.028]	[-0.712, 0.132]	4.257

Note. Final observation forecast MAE values across all timepoints for each model listed on the diagonal, beginning with the CPath reference, followed by the top performing learning model. Learning models are then ordered according to decreasing performance. Upper half of matrix contains numeric difference in metric, with percent difference in parenthesis, then one-sided bootstrapped proportional *p*-value; lower half of matrix contains the corresponding bootstrapped 95% confidence interval for the numeric difference. All differences are for the better performing model in the column relative to the poorer performing model in the row.

Table 75

Cross-Model Comparisons – Raw Bias of ADAS-Cog Score for Whole Subject Trajectories

Model	Reference	MERF	Boosted	Bagged	GLMM	FNN	CNN	LSTM
Reference	1.602	1.536 (-95.90%) <.001	1.495 (-93.30%) <.001	1.179 (-73.61%) <.001	1.175 (-73.35%) <.001	1.076 (-67.18%) <.001	0.527 (-32.88%) .001	0.244 (-15.20%) .075
MERF	[1.289, 1.797]	0.066	0.042 (-38.71%) .435	0.357 (-84.45%) .003	0.361 (-84.60%) <.001	0.460 (-87.50%) <.001	1.010 (-93.89%) <.001	1.293 (-95.16%) <.001
Boosted	[1.240, 1.736]	[-0.203, 0.279]	0.107	0.316 (-74.63%) .003	0.320 (-74.88%) .003	0.419 (-79.60%) <.001	0.968 (-90.03%) <.001	1.251 (-92.10%) <.001
Bagged	[0.897, 1.487]	[0.130, 0.590]	[0.079, 0.537]	0.423	0.004 (-0.97%) .492	0.103 (-19.58%) .219	0.653 (-60.68%) <.001	0.936 (-68.87%) <.001
GLMM	[0.906, 1.468]	[0.114, 0.622]	[0.079, 0.570]	[-0.246, 0.286]	0.427	0.099 (-18.80%) .248	0.648 (-60.29%) <.001	0.932 (-68.57%) <.001
FNN	[0.853, 1.297]	[0.209, 0.726]	[0.185, 0.671]	[-0.148, 0.385]	[-0.185, 0.374]	0.526	0.550 (-51.10%) <.001	0.833 (-61.29%) <.001
CNN	[0.219, 0.885]	[0.759, 1.272]	[0.734, 1.215]	[0.395, 0.946]	[0.361, 0.937]	[0.330, 0.806]	1.075	0.283 (-20.84%) .045
LSTM	[-0.089, 0.532]	[1.060, 1.539]	[1.008, 1.494]	[0.684, 1.219]	[0.636, 1.204]	[0.593, 1.055]	[-0.046, 0.605]	1.359

Note. Whole subject trajectory raw bias values across all timepoints for each model listed on the diagonal, beginning with the CPath reference, followed by the top performing learning model. Learning models are then ordered according to decreasing performance. Upper half of matrix contains numeric difference in metric, with percent difference in parenthesis, then one-sided bootstrapped proportional *p*-value; lower half of matrix contains the corresponding bootstrapped 95% confidence interval for the numeric difference. All differences are for the better performing model in the column relative to the poorer performing model in the row.

Table 76

Cross-Model Comparisons – Raw Bias of ADAS-Cog Score for Final Observation Forecasts

Model	Reference	Bagged	FNN	GLMM	CNN	LSTM	MERF	Boosted
Reference	-1.360	-1.152 (-84.71%) <.001	-1.792 (-68.31%) .005	-0.768 (-56.43%) .005	-0.292 (-21.47%) .227	-0.251 (-18.48%) .231	-0.209 (-15.37%) .249	-0.038 (-2.82%) .428
Bagged	[-1.775, -0.498]	-0.208	0.639 (-148.25%) .243	-0.385 (-64.91%) .138	-0.860 (-80.53%) .004	-0.901 (-81.24%) .005	-0.943 (-81.93%) .001	-1.114 (-84.26%) <.001
FNN	[-2.584, -1.096]	[-0.023, 1.307]	0.431	-1.024 (-172.73%) .320	-1.499 (-140.35%) .037	-1.540 (-138.87%) .028	-1.582 (-137.44%) .020	-1.753 (-132.61%) .006
GLMM	[-1.331, -0.216]	[-1.023, 0.271]	[-1.715, -0.287]	-0.593	-0.476 (-44.52%) .049	-0.516 (-46.55%) .033	-0.559 (-48.52%) .028	-0.729 (-55.16%) .014
CNN	[-1.062, 0.412]	[-1.474, -0.238]	[-2.255, -0.777]	[-1.041, 0.100]	-1.068	-0.041 (-3.67%) .430	-0.083 (-7.21%) .399	-0.254 (-19.19%) .234
LSTM	[-0.991, 0.432]	[-1.572, -0.200]	[-2.236, -0.876]	[-1.099, 0.046]	[-0.781, 0.660]	-1.109	-0.042 (-3.68%) .467	-0.213 (-16.11%) .271
MERF	[-0.796, 0.350]	[-1.608, -0.284]	[-2.286, -0.818]	[-1.158, 0.003]	[-0.804, 0.643]	[-0.743, 0.680]	-1.151	-0.171 (-12.91%) .291
Boosted	[-0.704, 0.582]	[-1.726, -0.454]	[-2.476, -1.050]	[-1.264, -0.087]	[-0.959, 0.418]	[-0.912, 0.457]	[-0.713, 0.377]	-1.322

Note. Final observation forecast raw bias values across all timepoints for each model listed on the diagonal, beginning with the CPath reference, followed by the top performing learning model. Learning models are then ordered according to decreasing performance. Upper half of matrix contains numeric difference in metric, with percent difference in parenthesis, then one-sided bootstrapped proportional *p*-value; lower half of matrix contains the corresponding bootstrapped 95% confidence interval for the numeric difference. All differences are for the better performing model in the column relative to the poorer performing model in the row.

Table 77

Cross-Model Comparisons – Absolute Value of the Bias of ADAS-Cog Score for Whole Subject Trajectories

Model	Reference	FNN	LSTM	MERF	Boosted	CNN	Bagged	GLMM
Reference	4.421	-1.591 (-35.99%) <.001	-1.275 (-28.84%) <.001	-1.258 (-28.46%) <.001	-1.194 (-27.01%) <.001	-0.994 (-22.48%) <.001	-0.864 (-19.55%) <.001	-0.748 (-16.92%) <.001
FNN	[-1.733, -1.429]	2.830	-0.316 (-10.04%) .001	-0.332 (-10.51%) <.001	-0.397 (-12.29%) <.001	-0.597 (-17.42%) <.001	-0.727 (-20.43%) <.001	-0.843 (-22.95%) <.001
LSTM	[-1.446, -1.092]	[-0.456, -0.140]	3.146	-0.017 (-0.53%) .354	-0.081 (-2.51%) .120	-0.281 (-8.21%) .002	-0.411 (-11.55%) <.001	-0.527 (-14.36%) <.001
MERF	[-1.449, -1.036]	[-0.478, -0.161]	[-0.203, 0.167]	3.162	-0.064 (-1.99%) .253	-0.265 (-7.72%) .008	-0.394 (-11.08%) <.001	-0.510 (-13.90%) <.001
Boosted	[-1.424, -1.016]	[-0.526, -0.220]	[-0.254, 0.090]	[-0.247, 0.128]	3.227	-0.201 (-5.85%) .017	-0.330 (-9.28%) <.001	-0.446 (-12.15%) <.001
CNN	[-1.230, -0.771]	[-0.736, -0.429]	[-0.465, -0.116]	[-0.435, -0.034]	[-0.400, -0.034]	3.427	-0.129 (-3.64%) .160	-0.246 (-6.69%) .013
Bagged	[-1.079, -0.604]	[-0.867, -0.565]	[-0.597, -0.227]	[-0.576, -0.204]	[-0.560, -0.169]	[-0.361, 0.086]	3.557	-0.116 (-3.17%) .170
GLMM	[-0.992, -0.532]	[-0.972, -0.680]	[-0.704, -0.344]	[-0.701, -0.288]	[-0.674, -0.281]	[-0.483, -0.028]	[-0.338, 0.158]	3.673

Note. Whole subject trajectory AVB values across all timepoints for each model listed on the diagonal, beginning with the CPath reference, followed by the top performing learning model. Learning models are then ordered according to decreasing performance. Upper half of matrix contains numeric difference in metric, with percent difference in parenthesis, then one-sided bootstrapped proportional *p*-value; lower half of matrix contains the corresponding bootstrapped 95% confidence interval for the numeric difference. All differences are for the better performing model in the column relative to the poorer performing model in the row.

Table 78

Cross-Model Comparisons – Absolute Value of the Bias of ADAS-Cog Score for Final Observation Forecasts

Model	Reference	CNN	LSTM	GLMM	MERF	Bagged	Boosted	FNN
Reference	4.799	-2.519 (-52.49%) <.001	-2.361 (-49.20%) <.001	-2.226 (-46.39%) <.001	-1.979 (-41.24%) <.001	-1.911 (-39.83%) <.001	-1.641 (-34.20%) <.001	-1.362 (-28.38%) <.001
CNN	[-2.857, -2.011]	2.280	-0.158 (-6.47%) .299	-0.293 (-11.37%) .110	-0.540 (-19.14%) .008	-0.607 (-21.04%) .005	-0.878 (-27.79%) <.001	-1.157 (-33.66%) <.001
LSTM	[-2.830, -1.990]	[-0.522, 0.347]	2.438	-0.135 (-5.24%) .241	-0.382 (-13.54%) .017	-0.450 (-15.57%) .007	-0.720 (-22.80%) <.001	-0.999 (-29.07%) <.001
GLMM	[-2.464, -1.839]	[-0.678, 0.212]	[-0.552, 0.222]	2.573	-0.247 (-8.76%) .102	-0.315 (-10.90%) .024	-0.585 (-18.53%) <.001	-0.864 (-25.14%) <.001
MERF	[-2.393, -1.649]	[-0.904, -0.096]	[-0.851, -0.025]	[-0.485, 0.058]	2.820	-0.068 (-2.35%) .405	-0.338 (-10.70%) .030	-0.617 (-17.95%) .001
Bagged	[-2.330, -1.677]	[-0.972, -0.103]	[-0.893, -0.093]	[-0.553, -0.010]	[-0.468, 0.290]	2.888	-0.270 (-8.56%) .009	-0.549 (-15.98%) <.001
Boosted	[-2.159, -1.413]	[-1.263, -0.430]	[-1.163, -0.363]	[-0.823, -0.249]	[-0.752, 0.019]	[-0.647, -0.036]	3.158	-0.279 (-8.12%) .019
FNN	[-1.962, -0.778]	[-1.524, -0.697]	[-1.442, -0.619]	[-1.102, -0.528]	[-1.031, -0.260]	[-0.926, -0.315]	[-0.824, -0.027]	3.437

Note. Final observation forecast AVB values across all timepoints for each model listed on the diagonal, beginning with the CPath reference, followed by the top performing learning model. Learning models are then ordered according to decreasing performance. Upper half of matrix contains numeric difference in metric, with percent difference in parenthesis, then one-sided bootstrapped proportional *p*-value; lower half of matrix contains the corresponding bootstrapped 95% confidence interval for the numeric difference. All differences are for the better performing model in the column relative to the poorer performing model in the row.

Table 79

Cross-Model Comparisons – Accuracy of CDR-Based Impairment for Whole Subject Trajectories

Model	Reference	CNN	LSTM	Boosted	Bagged	MERF	GLMM	FNN
Reference	0.711	0.126 (17.68%) <.001	0.124 (17.45%) <.001	0.114 (16.05%) <.001	0.107 (15.05%) <.001	0.105 (14.72%) <.001	0.093 (13.06%) <.001	0.090 (12.61%) <.001
CNN	[0.105, 0.147]	0.836	0.002 (0.20%) .439	0.012 (1.40%) .120	0.019 (2.28%) .048	0.021 (2.58%) .022	0.033 (4.09%) .002	0.036 (4.50%) <.001
LSTM	[0.103, 0.144]	[-0.018, 0.021]	0.835	0.010 (1.20%) .141	0.017 (2.08%) .053	0.019 (2.38%) .035	0.031 (3.89%) .002	0.034 (4.30%) .002
Boosted	[0.090, 0.135]	[-0.010, 0.031]	[-0.010, 0.030]	0.825	0.007 (0.87%) .235	0.009 (1.16%) .214	0.021 (2.65%) .023	0.024 (3.06%) .015
Bagged	[0.086, 0.126]	[-0.003, 0.039]	[-0.004, 0.038]	[-0.013, 0.029]	0.818	0.002 (0.29%) .422	0.014 (1.77%) .096	0.017 (2.17%) .062
MERF	[0.082, 0.125]	[0.001, 0.043]	[-0.003, 0.039]	[-0.013, 0.029]	[-0.018, 0.023]	0.815	0.012 (1.47%) .147	0.015 (1.87%) .084
GLMM	[0.071, 0.114]	[0.011, 0.055]	[0.011, 0.053]	[0.001, 0.042]	[-0.006, 0.035]	[-0.011, 0.034]	0.803	0.003 (0.39%) .402
FNN	[0.068, 0.111]	[0.015, 0.057]	[0.014, 0.055]	[0.003, 0.044]	[-0.005, 0.039]	[-0.006, 0.036]	[-0.020, 0.024]	0.800

Note. Whole subject trajectory accuracy values across all timepoints for each model listed on the diagonal, beginning with the logistic model reference, followed by the top performing learning model. Learning models are then ordered according to decreasing performance. Upper half of matrix contains numeric difference in metric, with percent difference in parenthesis, then one-sided bootstrapped proportional *p*-value; lower half of matrix contains the corresponding bootstrapped 95% confidence interval for the numeric difference. All differences are for the better performing model in the column relative to the poorer performing model in the row.

Table 80

Cross-Model Comparisons – Accuracy of CDR-Based Impairment for Final Observation Forecasts

Model	Reference	LSTM	MERF	GLMM	Bagged	CNN	Boosted	FNN
Reference	0.730	0.242 (33.10%) <.001	0.237 (32.46%) <.001	0.237 (32.46%) <.001	0.232 (31.82%) <.001	0.218 (29.90%) <.001	0.204 (27.98%) <.001	0.102 (13.91%) <.001
LSTM	[0.221, 0.261]	0.972	0.005 (0.48%) .242	0.005 (0.48%) .254	0.009 (0.97%) .162	0.023 (2.46%) .020	0.037 (4.00%) <.001	0.140 (16.85%) <.001
MERF	[0.209, 0.256]	[-0.016, 0.024]	0.967	0.000 (0.00%) .549	0.005 (0.49%) .432	0.019 (1.97%) .090	0.033 (3.50%) .012	0.136 (16.29%) <.001
GLMM	[0.214, 0.260]	[-0.016, 0.024]	[-0.028, 0.023]	0.967	0.005 (0.49%) .406	0.019 (1.97%) .100	0.033 (3.50%) .008	0.136 (16.29%) <.001
Bagged	[0.204, 0.256]	[-0.012, 0.028]	[-0.019, 0.023]	[-0.019, 0.023]	0.963	0.014 (1.48%) .186	0.028 (3.00%) .027	0.131 (15.73%) <.001
CNN	[0.194, 0.248]	[0.003, 0.043]	[-0.009, 0.042]	[-0.009, 0.042]	[-0.014, 0.037]	0.949	0.014 (1.50%) .144	0.117 (14.04%) <.001
Boosted	[0.172, 0.232]	[0.017, 0.057]	[0.005, 0.056]	[0.009, 0.051]	[0.000, 0.051]	[-0.015, 0.039]	0.935	0.103 (12.36%) <.001
FNN	[0.045, 0.148]	[0.119, 0.159]	[0.112, 0.159]	[0.107, 0.159]	[0.103, 0.154]	[0.088, 0.146]	[0.070, 0.131]	0.832

Note. Final observation forecast accuracy values across all timepoints for each model listed on the diagonal, beginning with the logistic model reference, followed by the top performing learning model. Learning models are then ordered according to decreasing performance. Upper half of matrix contains numeric difference in metric, with percent difference in parenthesis, then one-sided bootstrapped proportional *p*-value; lower half of matrix contains the corresponding bootstrapped 95% confidence interval for the numeric difference. All differences are for the better performing model in the column relative to the poorer performing model in the row.

Table 81

Cross-Model Comparisons – Precision of CDR-Based Impairment for Whole Subject Trajectories

Model	Reference	FNN	GLMM	Boosted	CNN	Bagged	MERF	LSTM
Reference	0.958	0.014 (1.48%) .010	0.010 (1.00%) .070	0.007 (0.76%) .134	0.005 (0.53%) .220	0.002 (0.25%) .364	0.002 (0.23%) .375	0.000 (0.05%) .479
FNN	[0.002, 0.025]	0.972	0.005 (0.47%) .214	0.007 (0.72%) .112	0.009 (0.94%) .064	0.012 (1.22%) .027	0.012 (1.24%) .039	0.014 (1.43%) .019
GLMM	[-0.003, 0.021]	[-0.008, 0.016]	0.967	0.002 (0.24%) .370	0.004 (0.46%) .225	0.007 (0.75%) .131	0.007 (0.76%) .113	0.009 (0.95%) .090
Boosted	[-0.007, 0.019]	[-0.005, 0.018]	[-0.011, 0.014]	0.965	0.002 (0.22%) .358	0.005 (0.51%) .215	0.005 (0.52%) .232	0.007 (0.71%) .146
CNN	[-0.009, 0.018]	[-0.003, 0.021]	[-0.008, 0.016]	[-0.010, 0.014]	0.963	0.003 (0.28%) .311	0.003 (0.30%) .314	0.005 (0.48%) .228
Bagged	[-0.011, 0.015]	[-0.000, 0.023]	[-0.005, 0.020]	[-0.008, 0.017]	[-0.011, 0.015]	0.960	0.000 (0.01%) .480	0.002 (0.20%) .371
MERF	[-0.011, 0.015]	[-0.001, 0.022]	[-0.006, 0.019]	[-0.008, 0.017]	[-0.010, 0.016]	[-0.015, 0.014]	0.960	0.002 (0.18%) .376
LSTM	[-0.013, 0.013]	[0.000, 0.025]	[-0.004, 0.021]	[-0.006, 0.018]	[-0.009, 0.018]	[-0.012, 0.015]	[-0.012, 0.014]	0.958

Note. Whole subject trajectory precision values across all timepoints for each model listed on the diagonal, beginning with the logistic model reference, followed by the top performing learning model. Learning models are then ordered according to decreasing performance. Upper half of matrix contains numeric difference in metric, with percent difference in parenthesis, then one-sided bootstrapped proportional *p*-value; lower half of matrix contains the corresponding bootstrapped 95% confidence interval for the numeric difference. All differences are for the better performing model in the column relative to the poorer performing model in the row.

Table 82

Cross-Model Comparisons – Precision of CDR-Based Impairment for Final Observation Forecasts

Model	Reference	CNN	GLMM	Bagged	LSTM	MERF	Boosted	FNN
Reference	0.977	0.018 (1.80%) .002	0.012 (1.24%) .068	0.012 (1.23%) .062	0.007 (0.68%) .162	0.007 (0.67%) .188	0.006 (0.60%) .236	0.003 (0.34%) .333
CNN	[0.005, 0.023]	0.994	0.005 (0.55%) .106	0.006 (0.56%) .121	0.011 (1.11%) .042	0.011 (1.12%) .049	0.012 (1.19%) .022	0.014 (1.45%) .017
GLMM	[-0.006, 0.023]	[-0.005, 0.011]	0.989	0.000 (0.01%) .459	0.005 (0.55%) .276	0.005 (0.56%) .274	0.006 (0.63%) .167	0.009 (0.90%) .141
Bagged	[-0.005, 0.023]	[-0.006, 0.011]	[-0.017, 0.011]	0.989	0.005 (0.54%) .274	0.005 (0.55%) .262	0.006 (0.62%) .189	0.009 (0.89%) .132
LSTM	[-0.012, 0.023]	[-0.001, 0.017]	[-0.013, 0.017]	[-0.012, 0.017]	0.983	0.000 (0.01%) .361	0.001 (0.08%) .350	0.003 (0.34%) .354
MERF	[-0.016, 0.023]	[-0.001, 0.017]	[-0.013, 0.017]	[-0.013, 0.017]	[-0.020, 0.017]	0.983	0.001 (0.07%) .384	0.003 (0.33%) .336
Boosted	[-0.015, 0.023]	[0.000, 0.018]	[-0.012, 0.018]	[-0.012, 0.018]	[-0.020, 0.018]	[-0.021, 0.018]	0.982	0.003 (0.26%) .355
FNN	[-0.022, 0.023]	[0.003, 0.020]	[-0.009, 0.020]	[-0.008, 0.020]	[-0.017, 0.020]	[-0.018, 0.020]	[-0.017, 0.020]	0.980

Note. Final observation forecast precision values across all timepoints for each model listed on the diagonal, beginning with the logistic model reference, followed by the top performing learning model. Learning models are then ordered according to decreasing performance. Upper half of matrix contains numeric difference in metric, with percent difference in parenthesis, then one-sided bootstrapped proportional *p*-value; lower half of matrix contains the corresponding bootstrapped 95% confidence interval for the numeric difference. All differences are for the better performing model in the column relative to the poorer performing model in the row.

Table 83

Cross-Model Comparisons – Recall of CDR-Based Impairment for Whole Subject Trajectories

Model	Reference	LSTM	CNN	Boosted	Bagged	MERF	GLMM	FNN
Reference	0.661	0.164 (24.76%) <.001	0.161 (24.44%) <.001	0.146 (22.03%) <.001	0.141 (21.27%) <.001	0.138 (20.82%) <.001	0.115 (17.47%) <.001	0.107 (16.26%) <.001
LSTM	[0.139, 0.188]	0.824	0.002 (0.26%) .430	0.018 (2.23%) .069	0.023 (2.87%) .023	0.026 (3.26%) .019	0.048 (6.20%) <.001	0.056 (7.31%) <.001
CNN	[0.137, 0.186]	[-0.021, 0.026]	0.822	0.016 (1.97%) .108	0.021 (2.61%) .043	0.024 (3.00%) .023	0.046 (5.93%) <.001	0.054 (7.03%) <.001
Boosted	[0.118, 0.169]	[-0.007, 0.041]	[-0.009, 0.038]	0.806	0.005 (0.63%) .339	0.008 (1.01%) .265	0.030 (3.88%) .007	0.038 (4.97%) <.001
Bagged	[0.115, 0.163]	[0.000, 0.046]	[-0.004, 0.045]	[-0.018, 0.031]	0.801	0.003 (0.38%) .407	0.025 (3.23%) .021	0.033 (4.31%) .006
MERF	[0.111, 0.161]	[0.001, 0.051]	[0.001, 0.050]	[-0.017, 0.032]	[-0.021, 0.027]	0.798	0.022 (2.85%) .047	0.030 (3.92%) .010
GLMM	[0.089, 0.141]	[0.023, 0.072]	[0.021, 0.070]	[0.006, 0.054]	[0.001, 0.050]	[-0.003, 0.048]	0.776	0.008 (1.05%) .272
FNN	[0.083, 0.133]	[0.032, 0.081]	[0.028, 0.078]	[0.014, 0.062]	[0.008, 0.057]	[0.005, 0.053]	[-0.020, 0.034]	0.768

Note. Whole subject trajectory recall values across all timepoints for each model listed on the diagonal, beginning with the logistic model reference, followed by the top performing learning model. Learning models are then ordered according to decreasing performance. Upper half of matrix contains numeric difference in metric, with percent difference in parenthesis, then one-sided bootstrapped proportional p-value; lower half of matrix contains the corresponding bootstrapped 95% confidence interval for the numeric difference. All differences are for the better performing model in the column relative to the poorer performing model in the row.

Table 84

Cross-Model Comparisons – Recall of CDR-Based Impairment for Final Observation Forecasts

Model	Reference	LSTM	MERF	GLMM	Bagged	CNN	Boosted	FNN
Reference	0.694	0.289 (41.59%) <.001	0.283 (40.78%) <.001	0.278 (39.98%) <.001	0.272 (39.17%) <.001	0.250 (35.96%) <.001	0.244 (35.15%) <.001	0.121 (17.45%) <.001
LSTM	[0.268, 0.306]	0.983	0.006 (0.57%) .176	0.011 (1.15%) .089	0.017 (1.73%) .034	0.039 (4.14%) <.001	0.045 (4.76%) <.001	0.168 (20.55%) <.001
MERF	[0.260, 0.300]	[-0.013, 0.022]	0.978	0.006 (0.57%) .314	0.011 (1.16%) .174	0.034 (3.55%) .002	0.039 (4.17%) .001	0.162 (19.86%) <.001
GLMM	[0.254, 0.300]	[-0.010, 0.028]	[-0.017, 0.023]	0.972	0.006 (0.58%) .323	0.028 (2.96%) .025	0.034 (3.57%) .008	0.156 (19.18%) <.001
Bagged	[0.243, 0.295]	[-0.003, 0.034]	[-0.012, 0.028]	[-0.022, 0.028]	0.966	0.022 (2.37%) .057	0.028 (2.98%) .017	0.151 (18.49%) <.001
CNN	[0.222, 0.284]	[0.022, 0.056]	[0.011, 0.050]	[-0.000, 0.050]	[-0.005, 0.045]	0.944	0.006 (0.60%) .286	0.128 (15.75%) <.001
Boosted	[0.206, 0.277]	[0.025, 0.061]	[0.014, 0.056]	[0.007, 0.056]	[0.002, 0.050]	[-0.027, 0.036]	0.939	0.123 (15.07%) <.001
FNN	[0.060, 0.177]	[0.148, 0.184]	[0.139, 0.179]	[0.128, 0.179]	[0.123, 0.173]	[0.095, 0.162]	[0.088, 0.153]	0.816

Note. Final observation forecast recall values across all timepoints for each model listed on the diagonal, beginning with the logistic model reference, followed by the top performing learning model. Learning models are then ordered according to decreasing performance. Upper half of matrix contains numeric difference in metric, with percent difference in parenthesis, then one-sided bootstrapped proportional *p*-value; lower half of matrix contains the corresponding bootstrapped 95% confidence interval for the numeric difference. All differences are for the better performing model in the column relative to the poorer performing model in the row.

Table 85

Cross-Model Comparisons – ROC AUC of CDR-Based Impairment for Whole Subject Trajectories

Model	Reference	Boosted	FNN	Bagged	MERF	GLMM	CNN	LSTM
Reference	0.841	0.075 (8.87%) <.001	0.066 (7.89%) <.001	0.066 (7.80%) <.001	0.061 (7.30%) <.001	0.051 (6.09%) <.001	0.013 (1.60%) <.001	0.007 (0.84%) <.001
Boosted	[0.059, 0.090]	0.916	0.008 (0.90%) .161	0.009 (0.99%) .116	0.013 (1.46%) .051	0.023 (2.62%) <.001	0.061 (7.15%) <.001	0.068 (7.96%) <.001
FNN	[0.051, 0.082]	[-0.009, 0.024]	0.908	0.001 (0.09%) .471	0.005 (0.55%) .245	0.015 (1.70%) .042	0.053 (6.19%) <.001	0.059 (7.00%) <.001
Bagged	[0.047, 0.082]	[-0.006, 0.025]	[-0.016, 0.017]	0.907	0.004 (0.46%) .326	0.014 (1.61%) .056	0.052 (6.09%) <.001	0.059 (6.90%) <.001
MERF	[0.043, 0.078]	[-0.003, 0.028]	[-0.010, 0.022]	[-0.013, 0.022]	0.903	0.010 (1.14%) .136	0.048 (5.61%) <.001	0.054 (6.41%) <.001
GLMM	[0.031, 0.068]	[0.007, 0.039]	[-0.002, 0.030]	[-0.003, 0.031]	[-0.009, 0.027]	0.892	0.038 (4.42%) <.001	0.044 (5.21%) <.001
CNN	[0.022, 0.066]	[0.044, 0.076]	[0.036, 0.068]	[0.034, 0.069]	[0.030, 0.065]	[0.018, 0.056]	0.855	0.006 (0.76%) .001
LSTM	[0.031, 0.072]	[0.050, 0.082]	[0.043, 0.076]	[0.040, 0.076]	[0.036, 0.071]	[0.024, 0.062]	[0.016, 0.058]	0.848

Note. Whole subject trajectory ROC AUC values across all timepoints for each model listed on the diagonal, beginning with the logistic model reference, followed by the top performing learning model. Learning models are then ordered according to decreasing performance. Upper half of matrix contains numeric difference in metric, with percent difference in parenthesis, then one-sided bootstrapped proportional *p*-value; lower half of matrix contains the corresponding bootstrapped 95% confidence interval for the numeric difference. All differences are for the better performing model in the column relative to the poorer performing model in the row.

Table 86

Cross-Model Comparisons – ROC AUC of CDR-Based Impairment for Final Observation Forecasts

Model	Reference	Bagged	MERF	GLMM	Boosted	CNN	LSTM	FNN
Reference	0.825	0.162 (19.66%) <.001	0.157 (19.01%) <.001	0.157 (19.01%) <.001	0.148 (17.90%) <.001	0.133 (16.15%) <.001	0.124 (15.06%) <.001	0.099 (12.00%) <.001
Bagged	[0.148, 0.173]	0.987	0.005 (0.55%) .217	0.005 (0.55%) .195	0.015 (1.49%) .028	0.029 (3.02%) .001	0.038 (4.00%) <.001	0.063 (6.84%) <.001
MERF	[0.132, 0.172]	[-0.011, 0.017]	0.981	0.000 (0.00%) .454	0.009 (0.94%) .182	0.024 (2.46%) .027	0.033 (3.43%) .005	0.058 (6.26%) <.001
GLMM	[0.133, 0.173]	[-0.011, 0.016]	[-0.025, 0.016]	0.981	0.009 (0.94%) .176	0.024 (2.46%) .025	0.033 (3.43%) .007	0.058 (6.26%) <.001
Boosted	[0.123, 0.168]	[-0.001, 0.026]	[-0.016, 0.025]	[-0.014, 0.025]	0.972	0.014 (1.51%) .121	0.023 (2.47%) .035	0.049 (5.27%) .002
CNN	[0.147, 0.174]	[0.012, 0.040]	[-0.001, 0.039]	[-0.000, 0.039]	[-0.011, 0.035]	0.958	0.009 (0.95%) <.001	0.034 (3.71%) <.001
LSTM	[0.133, 0.172]	[0.021, 0.049]	[0.007, 0.048]	[0.008, 0.049]	[-0.004, 0.043]	[0.024, 0.049]	0.949	0.025 (2.73%) <.001
FNN	[0.058, 0.131]	[0.048, 0.074]	[0.032, 0.074]	[0.035, 0.074]	[0.024, 0.068]	[0.049, 0.075]	[0.034, 0.071]	0.924

Note. Final observation forecast ROC AUC values across all timepoints for each model listed on the diagonal, beginning with the logistic model reference, followed by the top performing learning model. Learning models are then ordered according to decreasing performance. Upper half of matrix contains numeric difference in metric, with percent difference in parenthesis, then one-sided bootstrapped proportional *p*-value; lower half of matrix contains the corresponding bootstrapped 95% confidence interval for the numeric difference. All differences are for the better performing model in the column relative to the poorer performing model in the row.

Table 87

Subject-Specific Effects Performance Metrics Summary – Meta-Database Samplings

	RMSE			Mean AE			AV bias		
Subject-specific effects design	CPath reference	DN BR model	MERF model	CPath reference	DN BR model	MERF model	CPath reference	DN BR model	MERF model
Whole trajectories									
Population effects only	6.43 ± 0.196	6.42 ± 0.232	6.32 ± 0.239	4.92 ± 0.125	4.60 ± 0.137	4.64 ± 0.140	1.12 ± 0.312	0.60 ± 0.414	0.21 ± 0.259
Imputed subject effects	12.54 ± 0.124	7.79 ± 0.191	8.14 ± 0.179	9.19 ± 0.098	5.73 ± 0.117	6.21 ± 0.120	3.03 ± 0.308	0.22 ± 0.286	0.20 ± 0.212
Observation forecasts									
Population effects only	7.70 ± 0.348	7.78 ± 0.385	7.47 ± 0.380	5.79 ± 0.212	5.46 ± 0.223	5.43 ± 0.215	0.57 ± 0.482	1.02 ± 0.529	0.63 ± 0.277
Imputed subject effects	16.52 ± 0.195	8.93 ± 0.339	10.07 ± 0.291	12.20 ± 0.155	6.46 ± 0.204	7.81 ± 0.192	3.94 ± 0.497	0.58 ± 0.525	0.39 ± 0.454
Fitted subject effects		4.24 ± 0.238	4.44 ± 0.249		3.00 ± 0.129	3.15 ± 0.137		0.25 ± 0.256	0.27 ± 0.275

Note. Results presented as mean ± standard deviation for RMSE and MAE and median ± interquartile range for AVB.

Table 88

RMSE and AVB Subject-Specific Effects Cross-Model Comparisons – Meta-Database Samplings – Whole Subject Trajectories

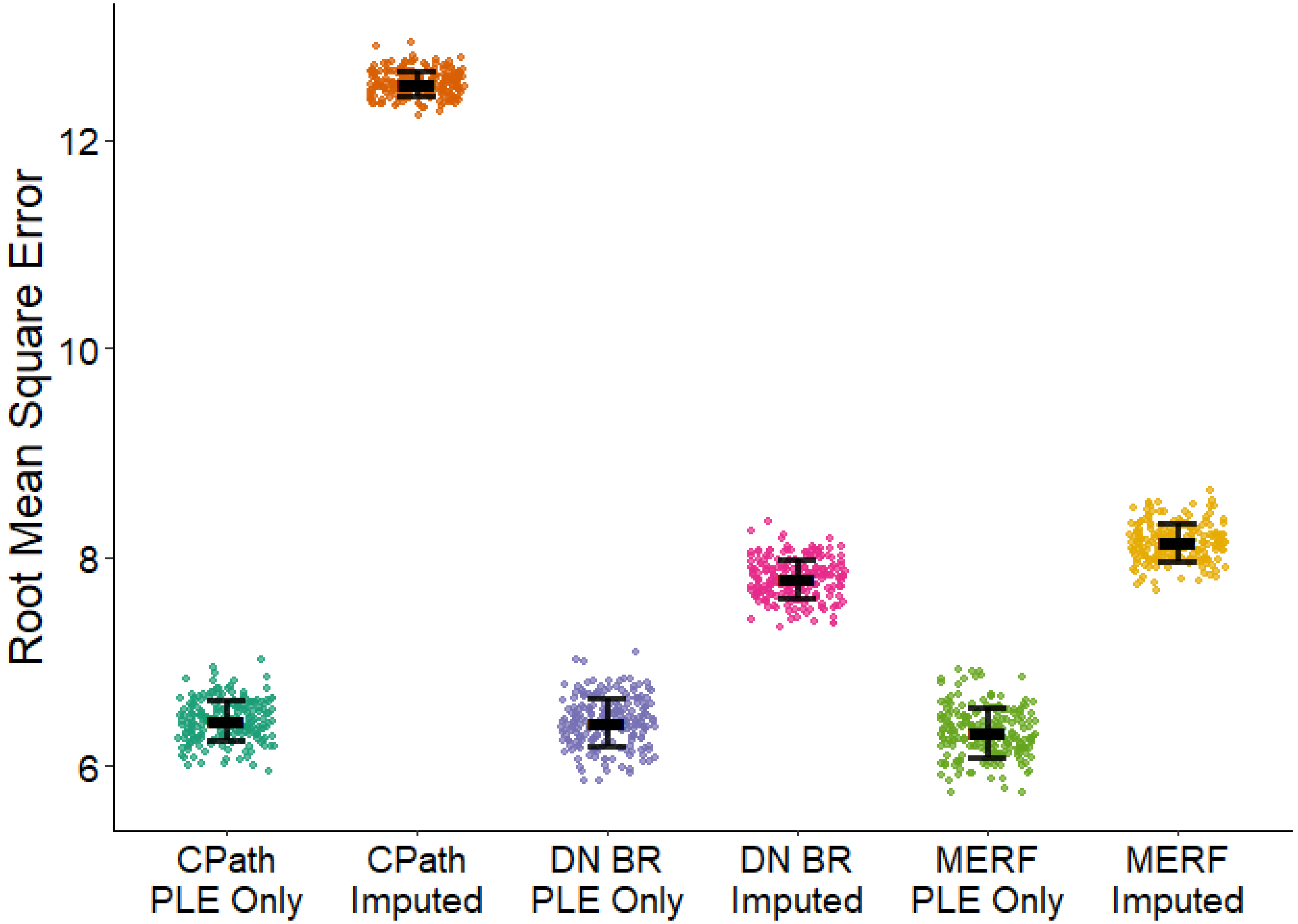
	CPath PLE only	CPath imputed	DN BR PLE only	DN BR imputed	MERF PLE only	MERF imputed
CPath PLE only	—	-171.0% [-201.4%, -153.5%] *	85.8% [52.2%, 141.6%]		434.1% [287.7%, 764.7%] *	
CPath imputed	95.0% [92.9%, 96.9%] *	—		1284.6% [825.4%, 1929.2%] *		1443.7% [1032.4%, 2052.6%] *
DN BR PLE only	-0.1% [-1.6%, 1.3%]		—	174.9% [68.4%, 292.9%]	187.4% [98.0%, 370.5%]	
DN BR imputed		-60.9% [-62.3%, -59.6%] *	21.3% [19.6%, 23.1%] *	—		11.5% [-38.1%, 84.9%]
MERF PLE only	-1.8% [-3.3%, -0.3%]		-1.6% [-3.3%, -0.1%]		—	6.6% [-57.9%, 65.9%]
MERF imputed		-54.0% [-55.2%, -52.9%] *		4.5% [3.5%, 5.6%]	28.8% [27.0%, 30.6%] *	—

Note. For whole subject trajectories, RMSE comparisons shown below the diagonal and AVB comparisons shown above. Results presented as percent differences in metrics of the row model relative to the column model (e.g. negative percentages indicate better predictive performance for the model design listed in the row relative to the column) along with bootstrapped 95% confidence interval for the percent difference. Tests not associated by either model type or subject-specific effects design were not evaluated with entries left blank.

* Comparisons where the better predicting model displayed improved metrics in at least 90% of the 200 meta-database samplings for whole subject trajectories.

Figure 79

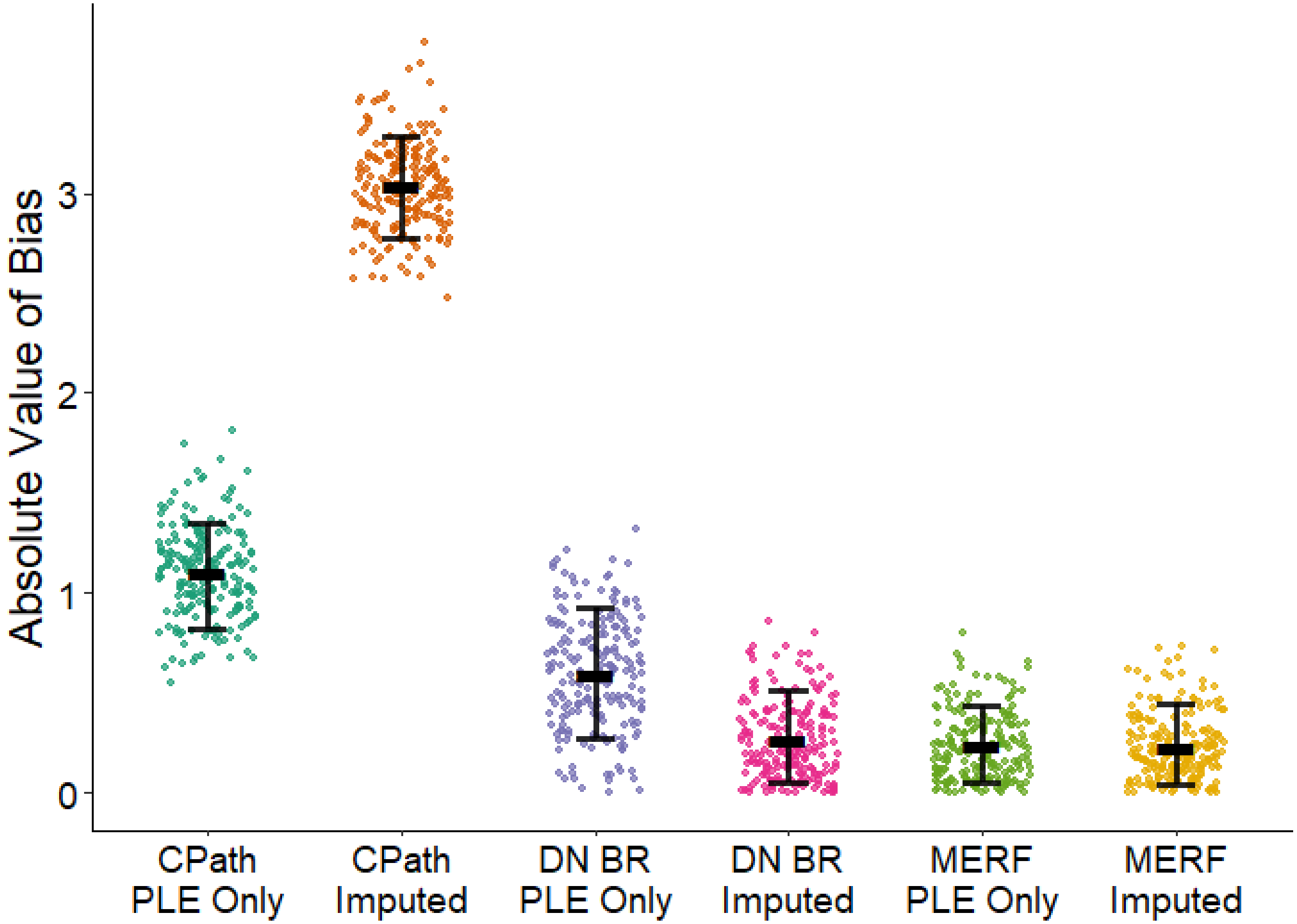
Root Mean Square Error on Subject-Specific Effects – Meta-Database Samplings – Whole Subject Trajectories



Note. Error bars centered on mean with standard deviation ranges.

Figure 80

Absolute Value of the Bias on Subject-Specific Effects – Meta-Database Samplings – Whole Subject Trajectories



Note. Error bars centered on median with interquartile ranges.

Table 89

RMSE and AVB Subject-Specific Effects Cross-Model Comparisons – Meta-Database Samplings – Final Observation Forecasts

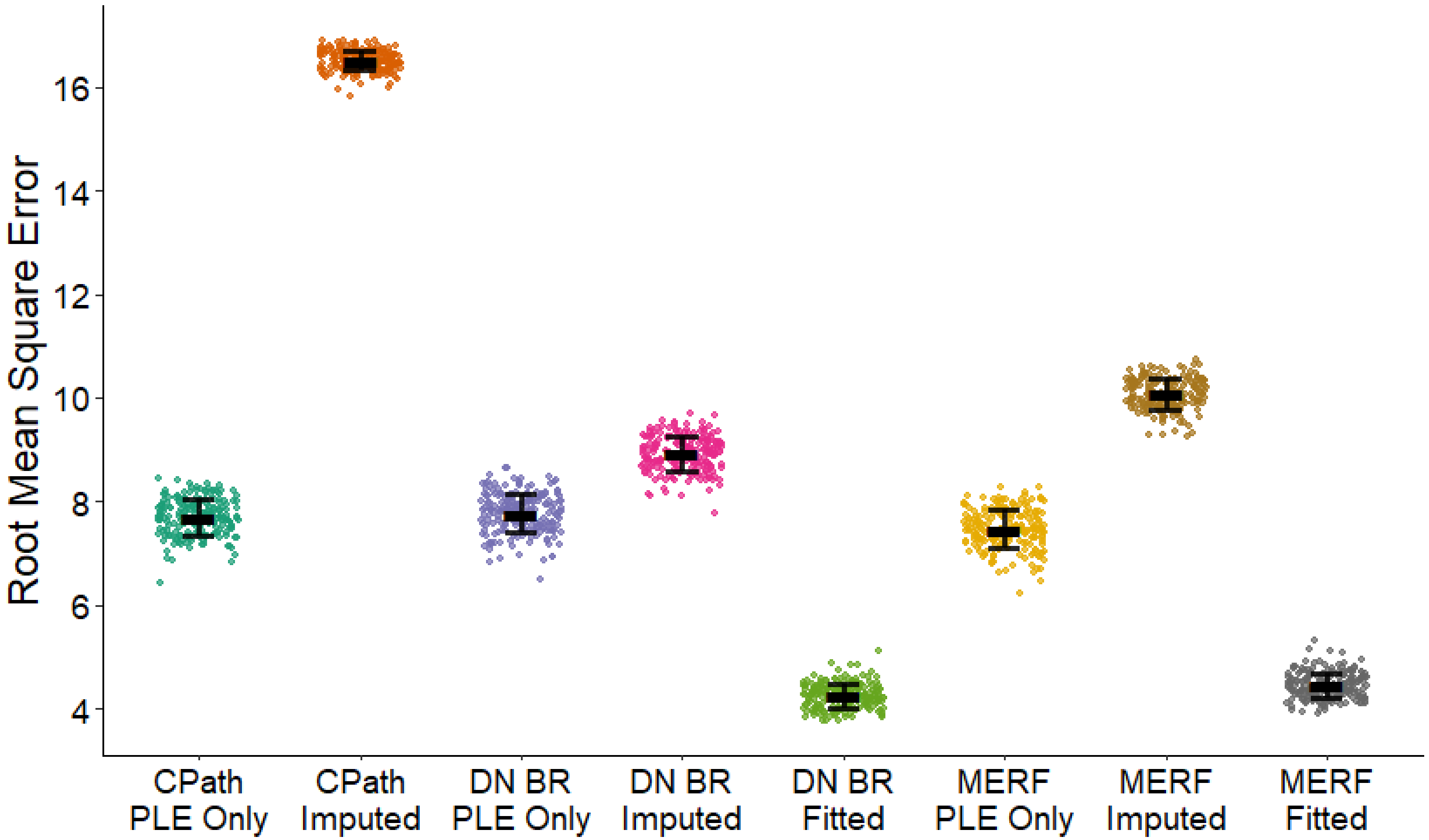
	CPath PLE only	CPath imputed	DN BR PLE only	DN BR imputed	DN BR fitted	MERF PLE only	MERF imputed	MERF fitted
CPath PLE only	—	-593.4% [-862.4%, -448.9%]*	-79.0% [-151.3%, -32.7%]			-11.8% [-59.0%, 17.3%]		
CPath imputed	114.7% [111.5%, 117.8%]*	—		575.9% [422.7%, 849.0%]*			899.3% [625.7%, 1462.8%]*	
DN BR PLE only	1.0% [-1.0%, 3.2%]		—	74.5% [25.7%, 143.2%]	305.2% [196.1%, 439.9%]*	60.1% [28.2%, 92.2%]		
DN BR imputed		-85.0% [-87.3%, -82.8%]*	14.9% [12.6%, 17.2%]*	—	132.3% [50.3%, 238.7%]*		47.8% [-3.5%, 152.3%]	
DN BR fitted			-83.4% [-87.5%, -79.3%]*	-110.7% [-114.8%, -106.3%]*	—			-8.5% [-52.8%, 35.9%]
MERF PLE only	-3.0% [-5.3%, -0.8%]		-4.1% [-6.4%, -1.9%]			—	61.1% [15.6%, 151.0%]	133.4% [79.0%, 213.8%]*
MERF imputed		-64.1% [-65.8%, -62.4%]*		12.7% [11.2%, 14.4%]*		34.8% [32.5%, 37.4%]*	—	44.9% [-6.3%, 124.1%]
MERF fitted					4.8% [2.2%, 7.4%]*	-68.1% [-72.0%, -64.2%]*	-126.6% [-130.7%, -122.4%]*	—

Note. For final observation forecasts, RMSE comparisons shown below the diagonal and AVB comparisons shown above. Results presented as percent differences in metrics of the row model relative to the column model (e.g. negative percentages indicate better predictive performance for the model design listed in the row relative to the column) along with bootstrapped 95% confidence interval for the percent difference. Tests not associated by either model type or subject-specific effects design were not evaluated with entries left blank.

[*] Comparisons where the better predicting model displayed improved metrics in at least 90% of the 200 meta-database samplings for final observation forecasts.

Figure 81

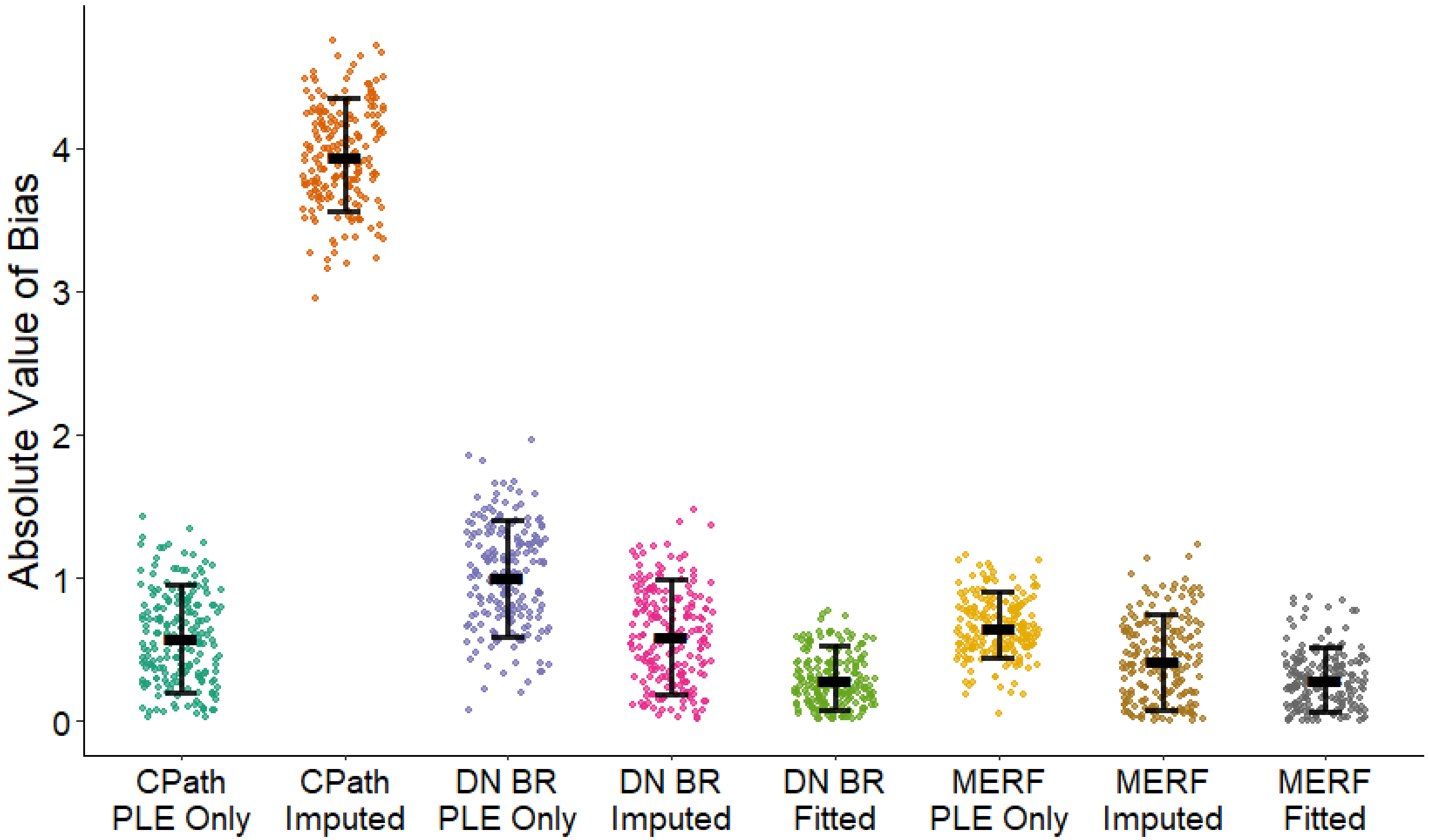
Root Mean Square Error on Subject-Specific Effects – Meta-Database Samplings – Final Observation Forecasts



Note. Error bars centered on mean with standard deviation ranges.

Figure 82

Absolute Value of the Bias on Subject-Specific Effects – Meta-Database Samplings – Final Observation Forecasts



Note. Error bars centered on median with interquartile ranges.

Table 90

Subject-Specific Effects Performance Metrics Summary – Synthetic Validation Cohorts

	RMSE			Mean AE			AV bias		
Subject-specific effects design	CPath reference	DN BR model	MERF model	CPath reference	DN BR model	MERF model	CPath reference	DN BR model	MERF model
Whole trajectories									
Population effects only	12.88 ± 1.552	11.80 ± 1.537	11.54 ± 1.441	9.54 ± 1.105	8.69 ± 1.071	8.84 ± 1.021	1.14 ± 1.677	2.24 ± 2.170	1.11 ± 1.387
Imputed subject effects	18.77 ± 0.904	15.75 ± 1.230	15.85 ± 1.373	13.89 ± 0.732	11.93 ± 0.942	12.25 ± 1.052	2.34 ± 2.120	1.08 ± 1.421	1.12 ± 1.423
Observation forecasts									
Population effects only	15.15 ± 2.101	13.04 ± 1.926	12.41 ± 1.849	11.28 ± 1.590	9.76 ± 1.432	9.47 ± 1.341	2.01 ± 2.435	2.23 ± 2.241	1.77 ± 2.043
Imputed subject effects	23.35 ± 1.267	17.03 ± 1.619	17.15 ± 1.685	17.88 ± 1.049	13.05 ± 1.263	13.51 ± 1.322	3.11 ± 2.829	1.19 ± 1.326	1.08 ± 1.353
Fitted subject effects		2.89 ± 1.548	2.65 ± 1.364		1.28 ± 0.492	1.19 ± 0.441		0.38 ± 0.470	0.44 ± 0.414

Note. Results presented as mean ± standard deviation for RMSE and MAE and median ± interquartile range for AVB.

Table 91

RMSE and AVB Subject-Specific Effects Cross-Model Comparisons – Synthetic Validation Cohorts – Whole Subject Trajectories

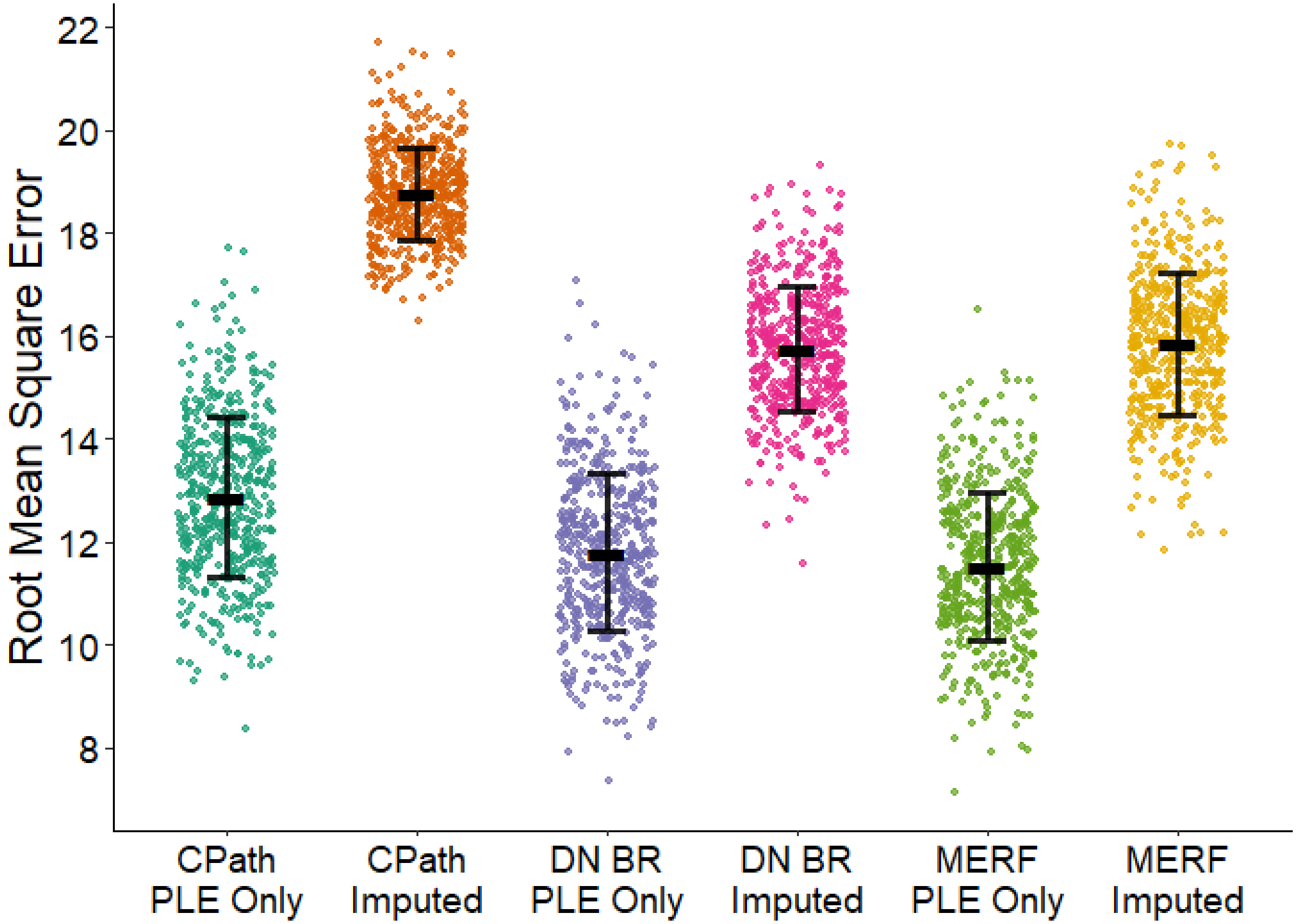
	CPath PLE only	CPath imputed	DN BR PLE only	DN BR imputed	MERF PLE only	MERF imputed
CPath PLE only	—	-106.3% [-175.6%, -48.2%]	-97.5% [-161.1%, -40.2%]		1.9% [-33.3%, 48.1%]	
CPath imputed	45.7% [42.2%, 49.5%] *	—		116.5% [70.4%, 194.1%]		108.3% [58.4%, 177.1%]
DN BR PLE only	-9.1% [-13.2%, -5.6%]		—	107.3% [57.8%, 185.7%]	101.2% [58.0%, 172.1%]	
DN BR imputed		-19.1% [-21.4%, -16.9%] *	33.5% [29.5%, 37.4%] *	—		-3.9% [-47.2%, 30.1%]
MERF PLE only	-11.6% [-15.5%, -7.8%]		-2.3% [-5.7%, 1.3%]		—	-0.9% [-43.2%, 35.8%]
MERF imputed		-18.4% [-20.8%, -16.1%] *		0.6% [-1.9%, 3.0%]	37.4% [33.0%, 41.9%] *	—

Note. For whole subject trajectories, RMSE comparisons shown below the diagonal and AVB comparisons shown above. Results presented as percent differences in metrics of the row model relative to the column model (e.g. negative percentages indicate better predictive performance for the model design listed in the row relative to the column) along with bootstrapped 95% confidence interval for the percent difference. Tests not associated by either model type or subject-specific effects design were not evaluated with entries left blank.

[*] Comparisons where the better predicting model displayed improved metrics in at least 90% of the 500 synthetic cohorts for whole subject trajectories.

Figure 83

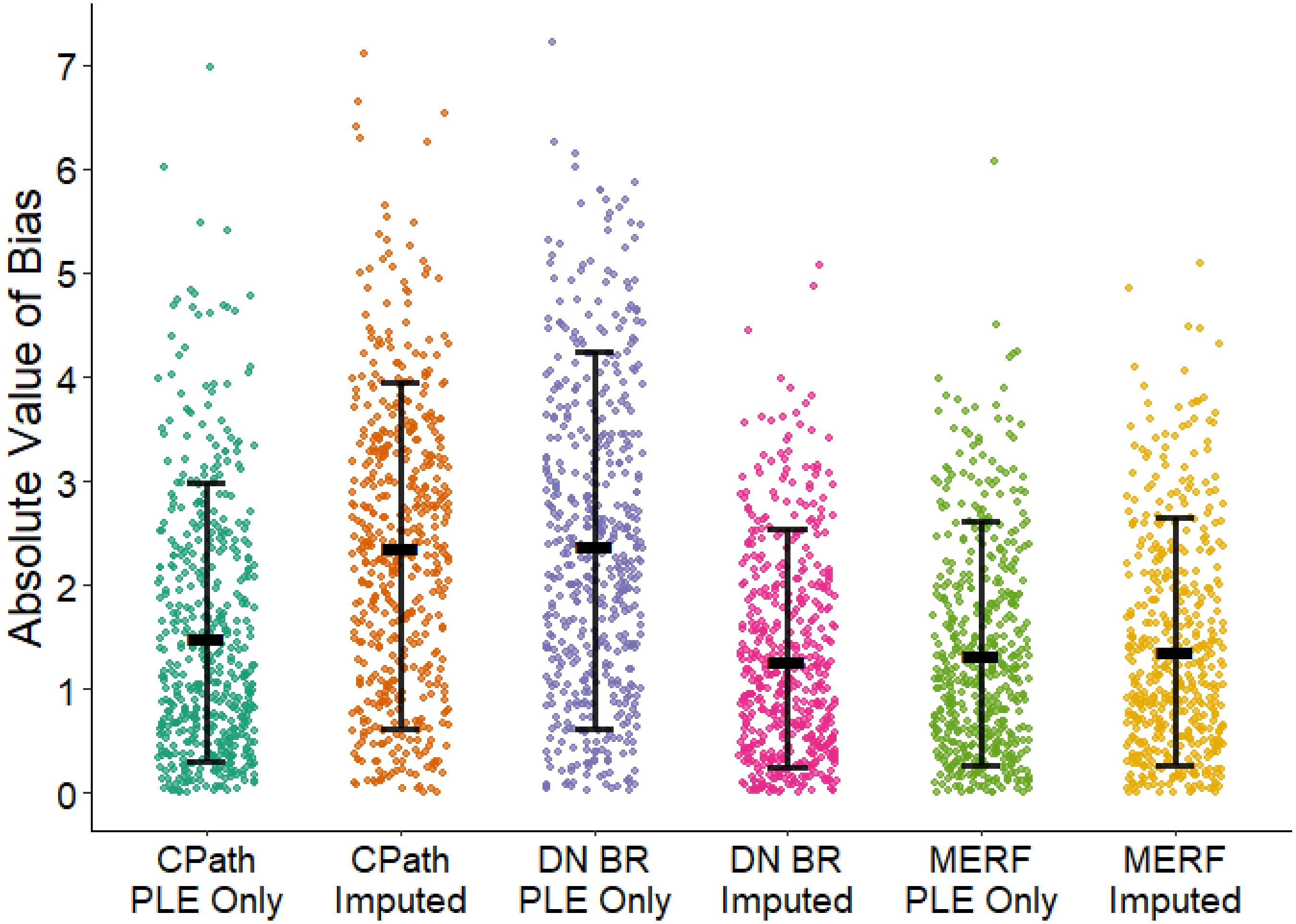
Root Mean Square Error on Subject-Specific Effects – Synthetic Validation Cohorts – Whole Subject Trajectories



Note. Error bars centered on mean with standard deviation ranges.

Figure 84

Absolute Value of the Bias on Subject-Specific Effects –Synthetic Validation Cohorts – Whole Subject Trajectories



Note. Error bars centered on median with interquartile ranges.

Table 92

RMSE and AVB Subject-Specific Effects Cross-Model Comparisons – Synthetic Validation Cohorts – Final Observation Forecasts

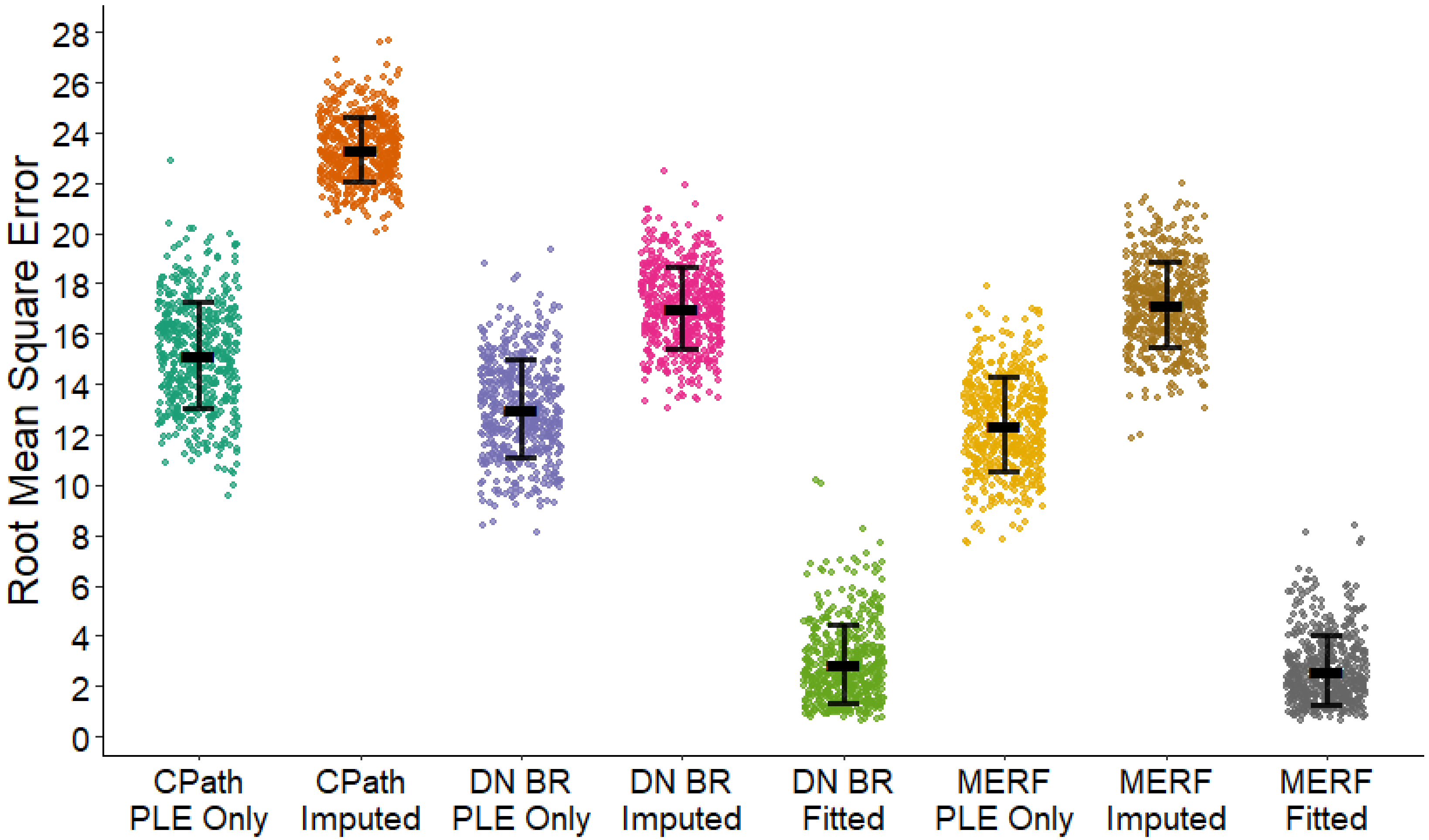
	CPath PLE only	CPath imputed	DN BR PLE only	DN BR imputed	DN BR fitted	MERF PLE only	MERF imputed	MERF fitted
CPath PLE only	—	-54.7% [-106.2%, -17.2%]	-11.2% [-51.9%, 19.7%]			13.6% [-18.8%, 55.0%]		
CPath imputed	54.1% [50.0%, 58.7%]*	—		161.2% [109.5%, 247.7%]			186.9% [121.5%, 281.2%]	
DN BR PLE only	-16.2% [-20.5%, -11.9%]		—	87.8% [43.8%, 158.9%]	480.1% [357.3%, 671.1%]*	26.3% [-1.7%, 69.6%]		
DN BR imputed		-37.1% [-40.1%, -34.2%]*	30.6% [26.4%, 35.1%]*	—	208.9% [133.2%, 315.5%]		9.8% [-25.3%, 47.3%]	
DN BR fitted			-350.6% [-399.7%, -303.9%]*	-488.5% [-561.0%, -427.8%]*	—			-14.0% [-46.3%, 15.8%]
MERF PLE only	-22.1% [-26.9%, -17.3%]		-5.1% [-9.4%, -0.6%]			—	63.3% [21.2%, 121.2%]	303.0% [212.6%, 418.8%]*
MERF imputed		-36.1% [-39.1%, -33.2%]*		0.7% [-2.3%, 3.4%]		38.2% [33.6%, 42.8%]*	—	146.8% [88.0%, 228.5%]
MERF fitted					-9.3% [-27.3%, 5.5%]	-369.0% [-422.1%, -325.0%]*	-548.2% [-622.5%, -488.7%]*	—

Note. For final observation forecasts, RMSE comparisons shown below the diagonal and AVB comparisons shown above. Results presented as percent differences in metrics of the row model relative to the column model (e.g. negative percentages indicate better predictive performance for the model design listed in the row relative to the column) along with bootstrapped 95% confidence interval for the percent difference. Tests not associated by either model type or subject-specific effects design were not evaluated with entries left blank.

[*] Comparisons where the better predicting model displayed improved metrics in at least 90% of the 500 synthetic cohorts for final observation forecasts.

Figure 85

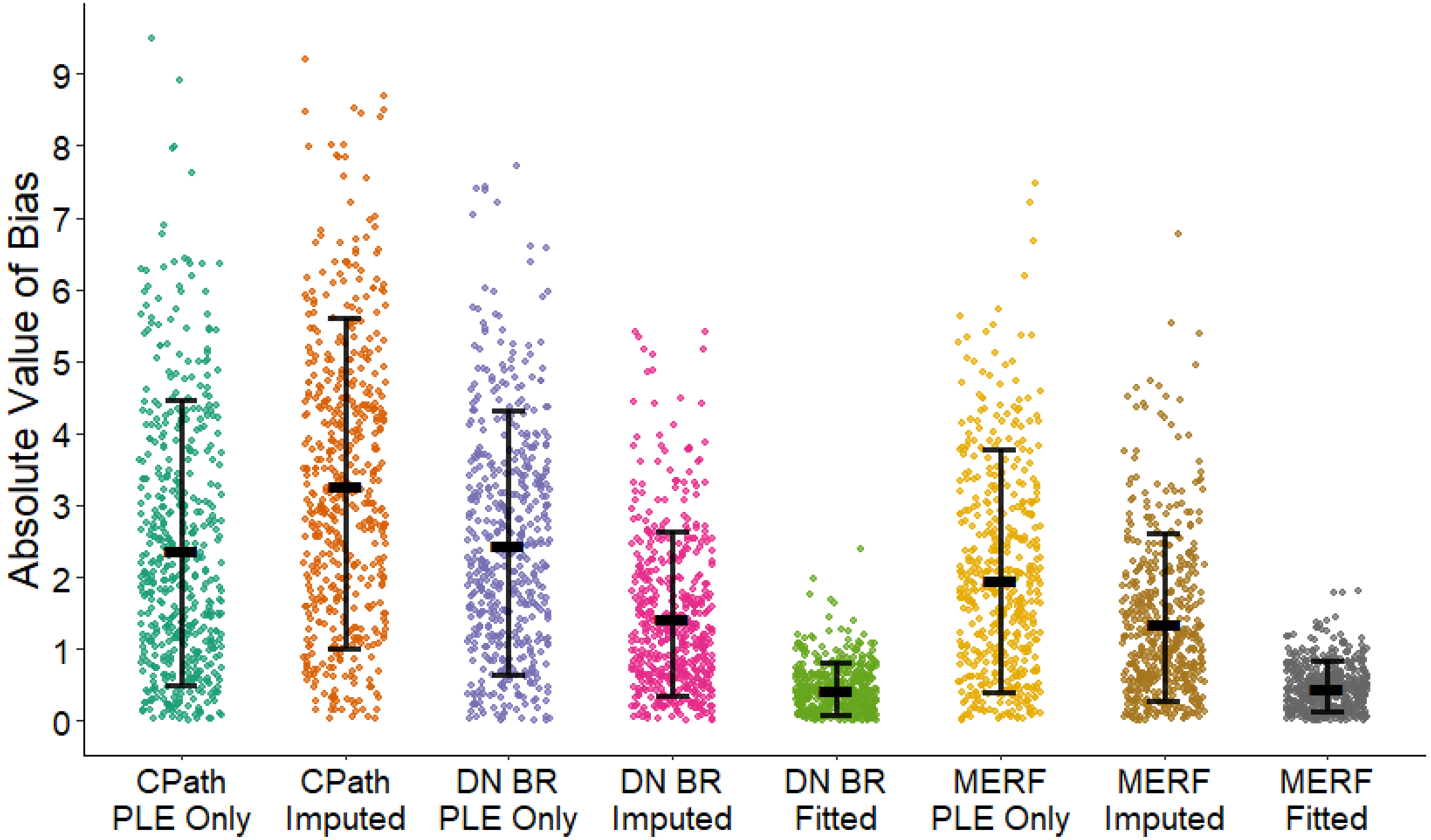
Root Mean Square Error on Subject-Specific Effects – Synthetic Validation Cohorts – Final Observation Forecasts



Note. Error bars centered on mean with standard deviation ranges.

Figure 86

Absolute Value of the Bias on Subject-Specific Effects – Synthetic Validation Cohorts – Final Observation Forecasts



Note. Error bars centered on median with interquartile ranges.

DISCUSSION

Overall Results

This dissertation is the first comprehensive evaluation of the predictive performance of longitudinal machine learning methods when applied to cognitive outcomes, considering ensemble supervised learning methods and novel adaptations of deep learning neural networks across multiple classes of responses with assessments for both generation of whole temporal profile trajectories and forecasting of future observations based on previously observed data. When compared against common inferential reference models with pre-specified parameterizations, all ML designs demonstrated improved performance under all evaluated conditions. Prediction errors and biases were reduced for predictions of ADAS-Cog scores while accuracy, sensitivity/recall and cut point agnostic ROC AUC measures all increased when classifying impairment status based on CDR scores. Furthermore, the models that made direct use of sequential data when forecasting future observations demonstrated exceptional gains in predictive performance compared to designs that only considered population-level covariates and ignored any possible subject-specific effects for both outcome types. Although all ML models showed improvement over their respective references, there was no one statistical learning design that was uniformly superior across all outcomes and types of predictions.

In addition to the ML evaluations, this study included the first assessment of the impact of subject-specific effects when predicting longitudinal change in cognitive decline as measured by the ADAS-Cog with comparisons considering a pre-parameterized model used for cohort generation, an analogous regression model built directly from the dataset, and an ensemble ML method using mixed-effects random forests. Imputation of subject-specific effects was associated with increases in prediction error compared to designs where these effects were suppressed but reductions in error were observed when using known fitted subject-specific effects. Bias was less consistent but was generally largest when only population-level effects were used and decreased when subject-specific effects were imputed and decreased further when fitted subject-specific effects were used directly during forecasting. However, for the pre-parameterized model, bias increased under imputation of subject-specific effects. Notably, although this increase was sizable within a real-world meta-database, it was significantly attenuated when tested under simulation with more generalized datasets and observed in only a fraction of the synthetic cohorts, implying generalizability of the reference model under imputation methods.

The Meta-Database of Alzheimer's Disease Studies

The meta-database proved to be an exceptionally potent data source for the evaluation aspects of this study, providing a large and robust collection of participants with ample amounts of longitudinal measures for both the continuous ADAS-Cog outcome and classification of CDR impairment. Even after multiple processing and harmonization steps and extraction of multiple data subsets for building both the logistic mixed-effects reference model as well as the holdout testing sets, there were still 2555 unique subjects with

ADAS-Cog measurements and 2755 participants with CDR scores, with nearly 15,000 total timepoints for model building. Importantly, the meta-database consisted of several different studies from both interventional trials of the ADCS and observational data from the ADNI study which supplied an excellent and varied set of ADRD study participants. This has implications for practical implementations of these ML models since they should readily generalize to other ADRD research populations and better highlight the utility of longitudinal machine learning methods to these subsequent cohorts.

While there is little question about how well the meta-database worked for this study, there are some discussion points which should be raised. Although some data were culled during the various cleaning steps of the harmonization, there was no larger loss than the requirement of *APOE4* genotyping for allele counts. Nearly half of the 7071 participants in the preceding phase were dropped during this step and three of the 14 ADCS studies were excluded solely on this requirement alone as they had no genotyping component. However, over half of the participants who were dropped for lack of genotyping only had single timepoint observations, with many of them being screen failures or study exclusions. This makes conceptual sense since investigators would not be expected to genotype a screen failure and this expectation helps offset some of the potential concern as the primary goal of this dissertation was to evaluate the longitudinal capacity of the statistical learning methods.

An early stage of the harmonization set the maximum evaluation time at six years even though some participants were followed for as long as 12 years. This decision to set the maximum time was not made lightly and was given substantial thought during the study design. The primary rationale was the ADCS interventional studies were typically conducted for either two or three years. As a result, nearly all of the observations at these more distal times came from the ADNI observational study. Even setting the maximum time to six years still gave a larger study footprint to the ADNI dataset at later times and this would have been further compounded with a more extended evaluation window. In addition, it is important to remember that Alzheimer's disease naturally focuses on an elderly population. This leads to a counterintuitive scenario wherein participants that are followed in an ADRD study for that exceptional length of time will tend to be "super agers" who are healthier and less likely to exhibit cognitive impairment. Paradoxically, this means participants observed at these later time points will have lower ADAS-Cog scores and less likely to have impairment based on the CDR. While it would have been intriguing to test the performance of the ML models at even longer time lengths, to avoid any issues with the apparent improvement in cognitive ability at later times, which would not be anticipated in a natural population, and further overemphasis on the ADNI dataset, it was decided to shorten the evaluation window to six years. Although disappointing, this still yielded an excellent cohort for predicting cognitive ability.

Another aspect of the meta-database to mention is the selection of the covariate feature space used for the ML models. A key benefit of machine learning in general is its capacity to work with larger feature spaces, even those that would lead to overfitting with fewer observations than features. Statistically, there would be expected improvements in prediction over the reference models simply be expanding the feature space, which is why relative increases to the reference models as well as cross-model comparisons were done for aims 1 and 2. In addition, when explicitly considering the role of subject-specific effects in aim 3, the same feature space was used for all models, even the MERF

ML model. This was to specifically isolate the impact of subject-specific parameters on ADAS-Cog prediction without additional influence due to wider feature spaces. When selecting the covariates for inclusion, other covariates known to influence cognitive ability were included such as race/ethnicity, education, and use of anti-dementia medication. It was unfortunate the only time dependent covariates which could be included were MMSE score and some vitals known to also associate with impairment, but this was ultimately dependent on the innate covariate coverage of the meta-database. The vitals measures of weight and blood pressure should not be taken as indications of vascular impairment as there are more appropriate metrics such as Hachinski score for stroke. While this metric is part of the meta-database, it did not have the same coverage as other covariates and was generally only seen at screening visits and used as an exclusionary tool. These listed metrics were selected due to their known association specific with ADRD based impairment and while other ML methods may highlight wider covariate sets, such as in genomic analysis, this still provided an appropriate feature space for the primary research question.

A final aspect to reiterate is the general disposition of the meta-database, especially with respect to the binary cognitive impairment classification generated from CDR score. A goal of this dissertation was to use a larger real-world dataset beyond the commonly utilized ADNI data frequently seen in AD-related ML studies. This was accomplished by the inclusions of the ADCS clinical trials but led to a unique set of concerns. The clinical trial data, as mentioned, is limited in duration to only 2-3 years. In addition, CDR is often used as an inclusion criterion, leading to the observed imbalance favoring impairment which is discussed further below. The more critical aspect is CDR is expected to change relatively little over a short period of time, nor is it expected to improve if cognitive or especially functional impairment is already present. A consequence of this is that CDR score, and by extension the impairment classifier of this dissertation, should be expected to be relatively stable within a given individual: people may transition from unimpaired to impaired if disease has progressed enough, but it is highly unlikely someone will transition in the opposite direction. Combined with the imbalance favoring impairment, there should be relatively little shift in the classifier. This undoubtedly is manifested in results of the impairment forecasting of the sequential models: rather than leveraging the covariate feature spaces, they are most likely relying on their known subject-specific effects leading to results such as the dramatic decrease in false negatives. While this is a point which absolutely must be raised, it does not necessarily detract from the research questions at hand. A primary goal was cross-model prediction and distinguishing recall sensitivities, even if they are exceedingly large, is still quite feasible with these methods. In addition, there is value in a model to identify stability. In fact, it would be a greater concern if these sequential models were unable to accurately forecast a consistent classifier. Regardless, taking the raw performance metrics at their exceedingly high face values is an incorrect representation of the results (frankly, no classifier should perform this well), but it does not detract from goals and results of this study in evaluating these longitudinal models in a more complex and realistic AD dataset.

Reference Model Observations

One of the most critical aspects when designing this study was the establishment of reference models for longitudinal prediction of the two cognitive outcomes. Analytically, this provided a control condition that all ML models could be compared to which helped ground the ML models to a common baseline during evaluation. Importantly, this would also provide an implementation reference since many of these novel ML methods, especially with the adaptations to account for panel data, show extremely high demands with respect to both processing speed and memory footprints. Although these practical implementation criteria were not directly evaluated during the course of the study, *ad hoc* observations were noted with a consistent eye to how the ML models compared to these inferential references both predictively and practically. In fact, one of the most striking results of the reference model implementations was how well both the CPath regression model for ADAS-Cog and the logistic mixed-effects impairment classifier performed when predicting either whole subject trajectories or final observation forecasts. This is especially notable since both reference designs had parameterizations determined from outside the evaluated meta-database, with the CPath model developed from legacy datasets and literature values, and the logistic mixed-effects classifier built with a holdout set that was never otherwise used during the ML evaluation process. Despite this, these inferential models were admirable for the predictive capacity indicating their generalizability to predict longitudinal cognitive outcomes even outside their original specifications. However, although both inferential methods had appreciable performance, there was still substantial room for improvement by the ML methods.

The CPath predictions for ADAS-Cog worked well in terms of on average values as reflected by the acceptable overall bias and mean errors for the two types of predictions. Observing overall RMSE/MAE values of 6.82/5.28 for trajectories and 9.72/6.86 for forecasts along with similar AVB values of 4.4 and 4.8 are appreciable for a scale that can range from 0 to 70. However, the model struggled to predict more extreme ADAS-Cog scores which became most apparent when inspecting the scatterplots of prediction scores. Many scores which were observed in the upper 40-70 point range were rarely predicted at that level. However, it should also be noted the CPath model struggled to predict lower scores and almost never gave a prediction in the 0-7 point range, values which would be expected of more cognitively intact individuals who were also known to be present in the meta-database. Thus, even though the scores were consistent for the CPath model, variances were much tighter and individual predictions had the potential to be vastly incorrect. A final point to reiterate is the CPath model performed better when predicting whole subject trajectories instead of forecasting final observations. This would be an important observation when predicting ADAS-Cog scores with the ML models, especially with the understanding that the use of the pre-parameterization precluded any direct leveraging of prior sequence data with reliance solely on the population-level demographic variables.

The logistic mixed-effects model similarly worked well overall in the meta-database and, unlike the CPath model, the decreased performance for forecasting was markedly attenuated. Accuracies for both types of predictions were greater than 0.70 with precisions greater than 0.95 indicating little concern with false positive misclassifications. However, both of these metrics, especially accuracy, must be carefully considered when used in imbalanced datasets such as the meta-database. As a trade-off to this large positive predictive capacity, there was a greater tendency of the logistic classifier towards false negative predictions wherein a participant would be misclassified as non-impaired with a CDR score of 0 while they were actually observed with an impairment classification from a larger CDR score. This was reflected most notably in the recall values with a 0.66 for the whole trajectories and 0.69 for forecasts. Another deficiency was how predictive performance generally decreased at later timepoints of three years or more with lower accuracy, precision, and recall, especially for trajectories. This was not generally observed with prediction of ADAS-Cog scores by the CPath model which was relatively consistent across timepoints and was more influenced by the presence of individual outlier values. A final aspect of these metrics to mention is their dependence on adequate cut point determination which must be considered for any score-based classifier. Taking the same data-driven approach of cut point optimization as described in the methods for all models, both inferential and learning based, certainly tempers this concern but this is where the utility of the cut point-agnostic ROC AUC becomes crucial. Despite the false negatives, a propensity generally observed with other evaluated models as well, the overall predictive performance of the logistic mixed-effects classifier was high with AUC values of 0.841 and 0.825 for whole subject trajectories and observation forecasts respectively. Within the overall context of this study, both reference models performed well, especially given their relatively limited covariate sets of age, MMSE, sex and *APOE4* count, but once again, clearly displayed room for improvement in their predictive ability for the cognitive outcomes of interest.

Original Implementation Goals for Supervised Methods

The initial proposal for aim 1 of this study was to evaluate a more comprehensive list of supervised machine learning models than the final series of ensemble tree methods. This included variations of regularized regression widely used in cross-sectional settings, most notably the ℓ2 penalized ridge regression, ℓ1 LASSO regression, and their combined elastic net formulation, as well as combination kernel methods of support vector machines. Given early citations, these methods appeared promising as additional components of more traditional supervised ML methods; however, practical application proved to be challenging. When attempting to implement the variety of regularized regression designs, many packages in R and libraries in Python which had first been published several years ago were no longer receiving active updates from the authors and were so outdated they no longer worked with the statistical software such as with `lmmlasso` and `glmmLasso` (Groll, 2017; Schelldorfer, 2011). Others had received more recent updates but were reliant on other packages and components which in turn had lost their own support and were no longer stable like `lmmen` (Sidi, 2020). Finally, some packages would work in their current software environments but were far too simplistic in their architecture for this study. This was the case with `ggmix` in R (Bhatnagar et al., 2021) and `GELMMnet` in Python (Schubert & Marks, 2017) which were originally developed for genomics studies and only allowed for single random effects components for intercept terms. While this was adequate for the panel data of their original implementations, they were insufficient for the complexities of the longitudinal designs of this dissertation.

For the multiple kernel SVMs, the original proposal was to have a standard Gaussian or radial basis function kernel account for the fixed effects components of the models while a second kernel would handle the panel nature of the data. This was inspired by the prior work of Luts et al. (Luts et al., 2012) who had demonstrated the potential of multikernel SVMs in both regression and classification of longitudinal data. A package was found for R which would handle multiple kernel SVMs which is still receiving support from the authors, `RMKL` (Wilson & Li, 2019). However, practical issues arose as pre-defined kernels of the package were found to be very limited and were highly reliant on specifications of the user. Implementation of an appropriate subject-specific kernel for panel data along with the proper selection and tuning of the multiple kernel design for the meta-database would constitute a notable research project on its own and, in the interests of giving the other models the credence and attention they deserved, it was decided to exclude the multiple kernel SVMs from this work.

While these early outcomes were disappointing, that is not to say they do no warrant investigation. Revisiting the regularized regression designs or implementing a multiple kernel SVM for longitudinal ADRD data may be worthwhile pursuits. However, it is telling that many of the citations are original publications from five to ten years ago with little subsequent follow-up. The machine learning field has seen many advances since then both with more refined ensemble methods like those of aim 1 and especially in the field of deep learning and neural networks as evaluated in aim 2. While there may still be utility in these other longitudinal supervised methods, the field as a whole may have simply decided on new avenues that could hold even greater potential and promise in the evaluation of repeated measures data. Ultimately, understanding and appreciating the direction the field is moving as a whole and adjusting accordingly is also an important facet for any research area. Finally, it is critical to reiterate that even if regularized regression and multiple kernel SVMs may not see much impact when applied in longitudinal or panel paradigms, they are still very widely used in statistical learning in cross-sectional applications. The `glmnet` package in R (Friedman et al., 2010) whose sole focus is regularized regression is one of the most robust statistical learning libraries available and is receiving constant use and support. In addition, SVMs are a critical component of the `scikit-learn` library in Python (Pedregosa et al., 2011) and are especially powerful in addressing high dimensional feature spaces. These ML models will continue to see use for quite some time and understanding their specific role and context, even if that does not involve longitudinal applications, is key to their utilization.

Classification of CDR-Based Impairment

Although CDR impairment for the meta-database in the context of within-subject stability has been discussed at length already, there are other aspects of this classification task to discuss in general terms. When conducting evaluations on a categorical outcome, it is critical to remember and reiterate two key points. The first aspect is how many of the metrics used in the evaluations of this study are reliant on a binary outcome as they arise from tabulations taken from a confusion matrix. Identifying false positives and false negatives requires explicit class assignments and without these there is no way to calculate metrics such as accuracy, precision/positive predictive value, or recall/sensitivity. Even other metrics that were not directly calculated here (e.g. the F1 score which is a linear transformation of precision and recall) have these same groundings in the confusion matrix. This bears mentioning since the outcomes of a classifier are almost never directly returned as the exact classes. Instead, they are generally score-based with a normalized value which is then reliant on informed decisions by the analyst or investigator to provide the corresponding classification assignments. Cut point optimization of these scores is a central aspect of categorization model building in its own right and can heavily influence the resulting calculations from a confusion matrix. Optimization may not even be the actual preference of an investigator if specific protection against type I or type II errors is a desired trait of the classifier. For the purposes of this dissertation, optimization based on data density and use of Youden's J statistic are more than sufficient and their consistent application across the reference, ensemble, and neural network methods does alleviate many of the concerns that may arise from cut point selection, but this process is still important to remember, nonetheless. In addition, the use of cut point agnostic evaluation metrics like the AUC of the ROC curve also help offset any complications that may arise from improper translation from a score to a binary class. These types of more comprehensive metrics are arguably even more valuable in properly evaluating a classification model.

The other point to mention is issues with class imbalance and the corresponding impact on the evaluation metrics, especially accuracy. In isolation, accuracy is actually a poor metric for model qualification in the presence of imbalance. After all, if one class is especially sparse, high accuracy can be achieved by simply always selecting the dominant class. Even in the absence of the observed within-subject stability, these imbalance effects could easily be a concern in this study on their own since the meta-database is known to have a preponderance of impairment with over 75% of the participants exhibiting CDR scores of 0.5 or greater at some point in their study. This imbalance could readily explain why precision in particular is so high for these models, even for trajectories. However, the evaluation pipeline itself is designed to address many of these concerns. First, the metrics themselves are never truly considered in isolation but only in relative changes. Thus, while accuracy may be high with the logistic mixed-effects reference model due to imbalance, it is the relative improvements in accuracy in the ML models that is the actual metric of interest. In addition, a variety of outcomes helps focus on the models themselves rather than the dataset. Changes in recall, for example, can help highlight a reduction in false negatives which would be a more likely occurrence in an imbalanced dataset, which could in turn represent a strength of the ML methods compared to other prediction designs. While these types of issues with classification model assessment are not of exceptional concern in the current study, they are worth bearing in mind whenever conducting any sort of evaluation of a categorization method.

A final practical consideration relates specifically to the observed imbalance of the meta-database. As mentioned, over 75% of the cohort was classified as having impairment according to the CDR. However, it must be noted CDR is still a clinical measure designed for a clinical setting. Furthermore, it is often used as a screening tool for exclusionary purposes in a research study. Issues with the veracity of the observed CDR score must at least be cited since there is the potential that CDR scores may be inflated for a variety of reasons. For example, the instrument itself may suffer due to inappropriate application by study personnel not sufficiently trained for such a clinically demanding tool or there is slight exaggeration between scores of 0 and 0.5 in the interest of meeting recruitment needs. If these cases are true, it may be that the models with relatively low recall are not predicting false negatives but are in fact accurately identifying these participants as cognitively unimpaired and it is the clinical ratings themselves that are at issue. This could easily not be the case, and this is certainly not a critique of the meta-database itself given the wide collection of studies. However, in the presence of the imbalanced disposition of CDR impairment this possibility is at least worth some commentary.

Ensemble Methods Observations

Two very noteworthy results were apparent during the assessments of aim 1. The first was how the supervised ensemble methods demonstrated improved predictive performance when compared to the reference models for both classes of outcomes and both types of predictions while being statistically significant in almost all evaluation metrics. The second was the substantial improvement when the longitudinal models were able to make direct use of prior sequence data while forecasting final observations when compared to *de novo* generation of whole subject trajectories. These results clearly demonstrate the utility and potential of the supervised ensemble methods as predictive models for the ADAS-Cog and impairment of CDR classification, especially with forecasting.

For the ADAS-Cog predictions, an interesting consideration is how well these models would predict extreme or outlier values, a noted issue with the CPath reference. Since RMSE is more sensitive to outliers compared to MAE, it would be anticipated that RMSE improvements would be more substantial if the specific utility of the ensemble methods was for predicting extreme scores. However, percent changes in RMSE and MAE were found to be similar for ML methods indicating uniform improvements in ADAS-Cog prediction across the full scores range. For trajectories, RMSE decreases ranged from 16.6% for the single GLMM tree to 27.7% for the MERF model while MAE improvements ranged from 16.2% to 28.2% for the same ensemble methods. Forecasts were not as close but still relatively similar with the top performer, the single GLMM tree, having RMSE and MAE improvements of 53.4% and 51.1% respectively while the weakest model, the boosted mixed-effects trees, observed improvements of 46.1% and

42.1%. Although not as well aligned as ADAS-Cog trajectories, there were fewer individual observations in the forecasting holdout set which may contribute to the increased variation. Regardless, the similarity in both RMSE and MAE highlight how well the ensemble methods reduce prediction error for the ADAS-Cog at all possible values, including extreme scores.

When considering improvements in bias, it is key to draw distinctions between raw bias, which can be negative, and absolute bias. Interestingly, raw bias indicated a tendency to overestimate ADAS-Cog scores when building trajectories but underestimate scores when forecasting final observations. This pattern was seen in both the CPath reference and the ensemble methods. However, while raw bias statistically improved for all ensemble methods when predicting trajectories, it only improved for the two GLMM methods but not the MERF or boosted mixed-effects models when forecasting. However, while bias can be informative to give indications or over or underestimation, as a metric it can be challenging to properly evaluate for predictive capacity, which is why absolute value of the bias was also considered. With AVB, the improvements in systematic error becomes clearer for the ensemble methods, with all models exhibiting significant improvements compared to the CPath reference for both types of predictions with improvements ranging from 16.9% to 28.5% for trajectories and 34.2% to 46.4% for forecasting of final observations. Combining the results of RMSE, MAE and AVB, we can easily see how the ensemble methods demonstrate superior predictive capacity by improving overall accuracy as well as the systematic components of the error.

For prediction of impairment by CDR scoring, the ensemble methods again demonstrated exceptional increases in performance compared to the reference logistic classifier.

Accuracy for trajectories increased for the ensemble models from 13.0% to 16.0% while forecast accuracy increased between 28.0% and 32.5% even with the known imbalance in impairment status. Although precision did increase numerically, none of the improvements were significant under bootstrapping although this is not especially surprising given the high positive predictive value already displayed by the reference model. Of greater importance was the reduction in false negative misclassification, a noted concern of the logistic model. Trajectories recall increased from 0.661 for the reference model to 0.776 for the single GLMM tree, the model with the smallest recall, to 0.806 for the top performing boosted mixed-effects trees. Forecasting improvements in recall were even more impressive with all models showing recall of 0.940 or greater. Outside explicit classification metrics, performance was still superior for the ML models with ROC AUC significantly increasing for both trajectories and forecasts. These improvements were moderate for trajectories, ranging from 6.1% for the single GLMM tree to 8.9% for the boosted mixed-effects model but were notably substantial when forecasting final observations with increases ranging from 17.9% to 20.0% with AUCs of 0.972 or greater. Even with the previously cited concerns when evaluating performance of these classifiers, the improvements for the ML models over the reference design were undeniable.

A final characteristic to note, as highlighted by the numeric values of the metrics, is just how much more powerful these models were when they were able to leverage previously observed data when forecasting final observations. The relative increases in metrics were substantial under forecasting, even when accounting for the minimal improvements in the logistic classifier and the actual worsening of predictive capacity for the CPath reference. Even allowing for these trends in the reference models is not enough to offset the vast improvements observed for the sequential ensemble methods when they were able to fully utilize their longitudinal capacity. This is perhaps best exemplified in the improvements in recall for CDR impairment as false negatives were a persistent concern for the logistic classifier and to some degree even the ML models when generating whole subject trajectories. Even when considering the discussed imbalanced and temporal stability concerns, forecasting sensitivity was at minimum 0.94 for the ensemble methods without any loss to precision with values of 0.98 or greater. This propensity for forecast prediction utilizing prior data, either using the more refined ADAS-Cog outcome or leveraging the knowledge of within-subject CDR stability, is one of the most telling aspects of this study and was a key observation leading into the evaluation of the neural network models of aim 2.

Neural Networks Observations

When first discussing the neural network designs, it is best to begin by highlighting how practically all the observations first presented in the ensemble methods of aim 1 also carried over when the sequential neural network adaptations were evaluated against the reference models in aim 2. First, just like the ensemble ML methods, there was substantial improvement in predictive performance for both classes of outcomes and both types of predictions. In addition, statistical significance was observed for performance improvements against the CPath regression and logistic classifier in all prediction metrics, with the exception of raw bias for observation forecasting of the ADAS-Cog and precision for classifying CDR impairment in either prediction type. Ostensibly, this would be for many of the same reasons such as increased variance in raw bias measure for ADAS-

Cog and the already high precision demonstrated by the logistic reference design when classifying CDR-based impairment. Finally, the improvements in predictive capacity for the ML models, especially the sequential 1D CNN and LSTM RNN, were especially apparent when applied to forecasting of final observations such that prior sequence information could be directly leveraged. Returning to the previously cited precision and recall when forecasting CDR impairment, the two sequential NN models proved the equal to the ensembles with recalls of at least 0.944 alongside exceptional positive predictive values of 0.983 at minimum. Considering the similarities in predictive ability between the supervised and neural network models, at least against the reference designs, what is more intriguing is how the NN methods differed amongst each other, especially when considering the non-sequential feed forward neural network and prediction of whole subject trajectories.

The standard feed-forward neural network was selected as an alternative reference model for the neural networks since it is by far the most ubiquitous deep learning model available. Every adaptation to neural networks builds off this original multi-layer perceptron design and, in many ways, it is the deep learning equivalent of the ordinary least-squares model in linear regression. This made it an excellent counterpoint to the other NN designs with their specific utility in longitudinal settings. What was most notable about the FNN is just how well it performed when generating whole subject trajectories. In fact, it was the best performing model when predicting ADAS-Cog, with the lowest RMSE, MAE and AVB values of all models, including the ensemble methods. It did lose some of this capacity in the classification design, with a tendency to misclassify false negatives leading to reductions in accuracy and recall although it did exhibit increased precision as well as an appreciable 0.908 ROC AUC for trajectories. Importantly, when compared to the whole subject trajectories of the 1D CNN and LSTM RNN sequential models, the non-sequential FNN displayed superior predictive capacity for the ADAS-Cog and better ROC AUC when classifying impairment.

At first, this seems to suggest the sequential design of the deep learning adaptations are not necessary when predicting cognitive performance; however, the FNN model did not have any way to directly leverage observed sequence data and, as a result, exhibited some of the lowest predictive ability when forecasting final observations. The RMSE for ADAS-Cog forecasts was comparatively poor for all neural networks, falling behind all the ensemble methods, and none were significantly different from any other. Even more striking is contrary to its exceptional performance for trajectories, the MAE and AVB under the FNN design were the worst of all evaluated ML models. This behavior was also seen when forecasting CDR impairment with the FNN having the lowest values for all classification metrics, a stark reversal of its performance for trajectories. Perhaps the most telling observation of the FNN predictions is how it mirrored the behavior of the reference models rather than the other ML paradigms: while the sequential ML methods almost universally saw great improvements in performance metrics with forecasting compared to generating whole subject trajectories, the non-sequential FNN, just like the reference models, saw only mild gains when classifying CDR impairment and actually worsened when forecasting ADAS-Cog scores.

There are two likely explanations for these patterns of the deep learning methods, one specifically related to the FNN and the other based on the sequential adaptations. First, the most obvious explanation is that feed-forward neural network only considered time as a population level covariate and, unlike the sequential ML methods, is unable to directly utilize prior sequence data. This could very easily result in the observed predictive shifts that mirrored those of the reference model and further emphasizes the utility of the true sequential methods when they are able to leverage prior data for forecasting. The other potential issue is not with the ability of the FNN to forecast observations but rather how the sequential 1D CNN and LSTM RNN models generated whole subject trajectories. As described in the methods, prediction using sequential neural networks is iterative in nature with each time step being added piecewise to populate the evaluation array. For the first timestep, the array slice corresponding to either outcome is completely empty and the requisite linear algebra is unable to be calculated. To account for this, a baseline FNN model was created to generate seed values that could populate the baseline timepoints and then allow the iterative sequential building of the response measure to proceed as normal. The standard FNN model does not have this requirement since the array slice, which is actually just a column vector, is calculated all at once. If the starting values for trajectory generation of the sequential networks are inappropriate, this could easily impact downstream calculations and errors in prediction could be propagated throughout an entire panel of subject data. In other words, it may not be that the FNN models were exceptionally powerful when generating trajectories, but rather the methods used to build these profiles for the sequential neural networks were comparatively poor. Alternative methods to build the outcome array slice for 1D CNN and LSTM RNN models may lead to improved trajectory prediction that could be on par with the FNN model and constitutes an interesting research direction in its own right. Regardless, one thing that cannot be denied is how well the CNN and RNN models performed when allowed to appropriately leverage previously observed data as exhibited by the same excellent performance in forecasting observed in the other sequential ML methods for both types of cognitive outcomes measures.

Cross-Model Comparisons for All Designs

While all the machine learning methods demonstrated substantial improvements over the inferential reference models, cross-model comparisons were also an important facet of this assessment. This helps further offset concerns such as the independent covariates of the ML models which may have artificially increased their predictive capacity simply by increasing their feature space used for fitting the outcomes. However, one of the most intriguing outcomes of this study is there was no blanket consensus as to the "best" model for either regression of ADAS-Cog or classification of CDR impairment. Additionally, aside from the observations of the neural network models previously described, there was no obvious model that outperformed within either trajectories or forecasting. Many of the models exhibited their own strengths and weaknesses and each could have their own utility depending on research goal and context.

Although unanimous agreement was not possible, some trends did arise. As mentioned, the FNN model performed especially well when predicting ADAS-Cog trajectories, outpacing all other models with respect to RMSE, MAE and AVB. This includes statistical significance over the MERF and boosted mixed-effects models which had similar metrics to each other but outperformed the two GLMM ensemble methods for ADAS-Cog trajectories. Trajectory prediction errors of the two GLMM models and the LSTM and CNN models clustered below these other models with the GLMM models in particular showing deficiencies in all metrics. However, when forecasting ADAS-Cog scores, the single GLMM tree improved substantially, with the lowest RMSE values with MAE values which were on par with the MERF errors, which had retained their low values first observed with the ADAS-Cog trajectories. Interestingly, the RMSE values for the LSTM and CNN models remained comparatively high but the MAE values were comparable to the MERF and single GLMM model. This pattern is intriguing for its implications on predictions of outlier values, suggesting the ensemble methods may particularly excel at finding these extreme values when compared to the neural networks which are more sensitive to outliers. Additionally, the AVB values for the sequential neural networks were the lowest for all models although no statistical difference was observed with the MERF or GLMM models, implying specific utility of the sequential ANN models for minimizing bias. Finally, although the boosted mixed-effects trees performed well with trajectories with the ADAS-Cog, they tended to lag behind the other models with respect to forecasting. The boosted trees displayed the highest AV bias and MAE values of all models, second only to the previously cited poor performance of the non-sequential FNN, and also had the highest RMSE values of all ensemble methods.

Some general cross-model trends were also apparent with respect to prediction of CDR impairment. Precision was known to be exceptionally high for all models and it was not especially surprising to observe that none of the models differed against one another, either when predicting whole subject trajectories or forecasting final observations. False negative protection as measured by recall indicated three clusters for trajectories with the sequential neural networks outperforming the ensemble methods while, as previously discussed, the FNN model demonstrated a relatively high false negative rate. Similar clustering for recall was seen with forecasting, although not as well defined simply because all the sequential models exhibited such vast improvements. However, the LSTM, MERF and two GLMM models were statistically indistinguishable with the CNN and boosted trees falling slightly behind while, again, the FNN model displayed exceptionally poor recall. Relative accuracy interestingly tended to favor the neural networks over the ensemble methods, most likely because of the increase in recall; however, statistical significance tended not to be seen simply because of the high accuracy already inherent in the dataset, ostensibly due to repeatedly cited meta-database imbalance. Removing consideration of cut points and focusing on the ROC AUC as a global metric was interesting since it favored the boosted mixed-effects and FNN for trajectories which indicates better cut point optimizations may have been possible for those two models in particular and alternatives to the data driven approach may have been valid. Forecasting comparisons are also valuable in the current context despite the impairment stability as all sequential models are expected to perform well. In these cases, AUC values clustered all of the ensemble methods ahead of the neural networks; however, this must again be tempered by the observation that the AUC differences for forecasting of the sequential models were comparisons of values of 0.98 and 0.96 which are both exceptionally high.

Whole Subject Trajectories and Final Observation Forecasting

As has been discussed several times, one of the most prominent results is how much predictive performance of the sequential ML models improved when they were able to make use of prior sequence data when forecasting final observations, whether due to improved fit as expected in ADAS-Cog or knowledge of classification stability as with CDR-based impairment. This behavior is best highlighted by omission, focusing on the pre-specified CPath regression and logistic classification reference models alongside the non-sequential FNN models. As mentioned, the classification capacity of the logistic reference model only saw minimal gains when forecasting while the CPath reference actually worsened with increases in prediction error and bias. However, by virtue of their parameterizations outside the training dataset, they were unable to make use of subject-specific effects and had to rely solely on the population level demographics. The FNN model exhibited similar behavior when forecasting for the same reason: all covariates were considered population-level effects, with cross-sectional time, and no subject-specific effects contributed to any part of the prediction. Meanwhile, forecasting with the sequential models provided predictions on meta-database subjects whose data had been part of their training dataset. As a result, subject-specific effects were known by these models, either in the form of fitted random effects for the ensemble methods that had mixed-effects components or the exact sequences of measures which comprised the outcome array slice for the sequential neural networks. However, these additional parameters were not available when the ML models generated whole subject trajectories since this testing set was comprised entirely of holdout data. As a result, they were unable to directly leverage their longitudinal capacity leading to decreases in predictive performance.

An important final point on CDR stability to mention is the direct comparison of forecasting for the two non-sequential methods in this dissertation: the logistic reference model and the feed-forward neural network. The stability of CDR within subjects cannot be the sole explanation of the exceptionally high classification metrics while forecasting for the sequential models. If this were the case, the FNN model would not have had such marked numeric and statistical improvements over the reference classifier since neither of these designs used any sort of subject-specific component. At least some of the improvement in forecast predictions of the impairment classifier must come from the use of the deep learning methodology over the inferential method, otherwise improvements in the FNN model would be much less apparent. Whether due to the expanded feature space, differences in estimation architecture, or some other rationale, there must be at least some aspect of machine learning which aids in prediction and forecasting the impairment classification beyond leveraging subject-specific effects.

Accordingly, one question that arises naturally from this is the exact role these subject-specific effects have on predicting cognitive outcomes. Is the prior knowledge the most significant contribution or is simply the inclusion of the subject-specific effects merely as additional parameters enough to improve predictive ability? This was investigated early in the study when deciding how to handle subject-specific effects for the reference models. Both inferential methods are generalized linear mixed-effects models and are able to utilize random effects to account for subject-specific panel data. Even though exact fitted effects were unknown, covariance parameters for intercepts and slopes were provided by the CPath model authors and were calculated *ad hoc* when the logistic classifier was first built. This left two potential solutions for handling the subject-specific effects: either suppress them entirely to rely solely on the population-level fixed effects or impute them from the model covariances as reasonable approximations. Pilot attempts at imputation yielded unexpected results: all metrics for both ADAS-Cog scoring and CDR impairment classification were markedly worse under imputation for both trajectories and forecasts. This was especially surprising since the CPath authors recommend the use of the provided covariances when generating synthetic cohorts for feasibility of clinical trials. It was ultimately decided for aims 1 and 2 to instead rely solely on population-level effects for the reference designs, but these results informed much of the direction of aim 3 to identify the particular role subject-specific effects play in prediction of cognitive outcomes using these methods.

Influence of Subject-Specific Effects

While aim 3 could have been more comprehensive, addressing several models as well as CDR impairment, it was decided to focus on the primary research question of subject-specific effects and their utility in response prediction. The within-subject stability of CDR in the meta-database made it a poor candidate for this type of query so the emphasis was given to ADAS-Cog. The CPath model was taken based on the previous observation of imputation worsening prediction combined with the cited recommendation of imputation by the package authors. The *ad hoc* beta-regression model was selected as it followed the same covariate specification as the CPath model with the benefit of being built directly from the datasets with updated parameter values. This not only evaluated the impact of the pre-specified parameterization of the CPath model but also allowed for the application of known and fitted subject-specific effects for utilization during forecasting for the DN BR model. The MERF model was selected as it was among the top performing ML models when predicting the ADAS-Cog although the covariate set was reduced to the CPath selection to avoid any undue influence in evaluation metrics that may have arisen due to the expansion of the feature space. Together, this provided an excellent framework to specifically address the question of subject-specific effects and their impact on prediction.

Most prevalent was the pervasive increase in RMSE and MAE across all models when subject specific effects were robustly imputed compared to utilization of only population-level effects, largely due to increases in variance of the ADAS Cog predictions. However, although error of the ADAS-Cog predictions increased under imputation, the predictions were, on average, almost always less biased. This is intuitive from a strict statistical perspective as subject-specific effects are still parameters themselves, and their inclusion should lead to reduced bias even with increased variance (Hastie et al., 2009). The exception was imputation for the CPath model which observed increases in AVB for both trajectories and forecasting. Additionally, the increases in RMSE were much larger for the CPath model under imputation when applied to the meta-database, doubling in magnitude while the *de novo* BR and MERF models only increased by 20-30%.

At first, this would seem to imply the recommendation of the CPath authors to impute subject-specific effects is ill advised. However, just as striking were the results when imputation was investigated under the generalized synthetic datasets. Prediction error and bias increased for all models, expected given the more variable datasets, but the relative increases under imputation were vastly different. RMSE and MAE shifts for the *de novo* models under imputation increased to 35-40% while the CPath model, instead of doubling as seen in the meta-database, only observed increases in prediction error of 45%. Although bias still increased for the CPath model in the synthetic datasets, the attenuation was even sharper. Trajectory increases went from 170% in the meta-database to 106% in the synthetic cohorts while increase in forecasting bias was only 55% under simulation compared to a nearly 600% increase in the real-world dataset. While bootstrapping still placed the RMSE and MAE errors of the CPath model ahead of the *de novo* designs, bias under imputation or in the absence of subject-specific effects was statistically indistinguishable among the three evaluated models. Together, these results imply good generalizability of the CPath model with imputation of subject-specific effects, supporting the recommendations of the authors for its appropriate use case of cohort generation.

Fitted Effects from Previously Observed Data

Two important results also arise from the behavior of the *de novo* BR model and MERF model built directly from the datasets. First, a previously cited key benefit of building directly from the data is the ability to directly leverage known and fitted subject-specific effects when forecasting future observations in subjects who were part of the parameterization process. For both *de novo* models, properly utilizing the subject-specific effects as known parameters led to ADAS-Cog predictions which far outperformed the models using only population level effects or robustly imputing individual-level effects. While not necessarily surprising, it highlights just how powerful known subject effects can be with forecasting as the reductions in error and bias in both the meta-database and synthetic cohorts were statistically significant compared to the designs that either suppressed or imputed subject-specific effects. However, it must be noted this advantage only applies when both prior data can be used and when model parameterizations are done *ad hoc*. These advantages are unfortunately not applicable when either generating whole subject trajectories or when predicting outcomes using pre-specified parameterizations like those of the CPath model.

The other outcome to note is the similarity in error and bias values between the *de novo* BR and MERF models within the various designs of subject-specific effects. A primary result of aims 1 and 2 is the improvement of the ML predictions compared to the inferential references; however, as previously mentioned, the ML models had the benefit of an expanded feature space with additional explanatory covariates. The MERF model of aim 3 did not have this benefit, using the same covariate set as the other models. Thus, it was interesting to see how indistinguishable the errors and biases were when compared to the DN BR model which could have been cast as an inferential reference model itself in the earlier aims. However, these same results may not have been observed had the *de novo* BR model been expanded to include the additional covariates like race and education instead of reducing the feature space of the MERF model as was done here. A strength of ML models in general is their ability to accommodate especially wide datasets while inferential methods can be subject to overfitting. Building the aim 3 *de novo* models with this in mind could have easily led to improved performance for the MERF model as seen in aim 1 and demonstrated its utility in ADAS-Cog prediction over inferential methods when using a wider feature space. However, the goal of this aspect of the study was to specifically investigate the subject specific effects on prediction, thus it was deemed important to match the population-level effects and individual-level designs across the models. This led to a focused reliance on the architecture of the CPath model when defining the *de novo* models and suggests further investigation of the role of feature space when comparing inferential and machine learning methods.

Clinical Relevance of Imputation

From a clinical perspective, the most important consideration of this facet of the dissertation is how these observed patterns in error and bias should direct model utilization in ADRD research. Ultimately, this is dependent on the goals of the investigator as different types of predictions and model designs have their own strengths and weaknesses and are best used in their appropriate usage scenarios. Statistically, increases in error and bias are generally viewed as detrimental but they can serve certain purposes within larger clinical study and trial design contexts. This comes into sharpest focus when considering generation of synthetic data when compared to direct prediction of outcomes for an actual AD patient.

Although the current study focused on ADAS-Cog prediction, the CPath model itself was originally developed to generate feasible cohorts to simulate studies in cognitive decline, most notably interventional clinical trials. This is a very different goal from explicitly predicting either trajectories of decline or forecasting future events and the CPath parameterizations were not originally meant for these prediction tasks. Instead, the goal was to develop a methodology which could generate a reasonable cohort an investigator could anticipate recruiting for feasibility purposes. Under this structure, emphasizing data generation, increased variance and reduced bias are in fact desirable as they will lead to expected ADAS Cog trajectories which, on average, will tend to be close to the ground truth for a population of interest. This suggests the inclusion of subject-specific effects which have been imputed will give more appropriate outcomes with greater variability and reduced overall bias.

In comparison, when attempting to predict outcomes for an actual AD study participant or clinical patient, it is far more desirable to have outcomes that are highly accurate with the lowest possible error, even if these predictions are not as unbiased as they could be. In these situations, relying solely on population-level effects leads to more accurate results rather than imparting additional variance by imputing a potential, but otherwise unknown, subject-specific effect. Even more preferred is making use of previously observed data in an already evaluated individual to calculate known and fitted subject-specific effects which give the most accurate predictions possible. Thus, for predictions at the individual level, researchers and clinicians are best served building their own models if possible and using previously observed data or, if relying on outside parameterizations, only using demographic and population-level effects.

A key point to mention is when attempting to predict ADAS-Cog scores in the meta-database, a collection of real-world data with certain aspects which have been discussed at length, the pre-parameterized CPath model showed large increases in both error and bias. As mentioned, this would initially suggest the model as a whole is poorly designed for either generation of synthetic cohorts, its intended purpose, or prediction. However, it is critical to reiterate this study made use of prediction as a framework, for which the CPath model is not necessarily well suited. Critically, when applied to the more generalized synthetic cohorts during simulation, the increases in error and bias were markedly smaller in magnitude and much more in line with the evaluation metrics observed in the two models built directly from the data. This suggests a specific issue of using the pre-specified parameterization with the meta-database in particular given its disposition specifically under a prediction framework. However, based on the simulation results, the

CPath model can be expected to generalize well to other cohorts like those generated during simulation. This makes its utilization, including the imputation of subject-specific effects as recommended by the authors, more than appropriate for its intended use of generating cohorts for clinical trial simulation and calculating expected on-average ADAS-Cog values.

Limitations and future directions

Several of the limitations of this study have been previously mentioned but provide many possible avenues for further investigation. The cohort make-up of the meta-database has been commented on repeatedly but bears a final mention. The utility of any model is very dependent on the nature of the data it is applied to and these ML methods are no exception. Issues such as persistence of the CDR-based classifier within subjects, the imbalance emphasizing impairment, the loss of subjects missing genotyping, these all contribute to the performance of the models. This may also be compounded by the mixture of clinical and observational data in the meta-database. This mixture may be detrimental to prediction as these two cohorts have several inherent differences and should not necessarily be considered to have arisen from the same source population. As such, enrichment methods, like focusing on ADCS or ADNI studies independently, may represent an additional source of rigor. Data selection in turn feeds into availability of covariates from a dataset, like *APOE4*, and the associated reference design structure comes into play. A model built *ad hoc* may be a better reference, as seen with the *de novo* BR model of aim 3, than a pre-specified parameterization such as the CPath model used as the common reference design. Of course, not every investigator has the data to build their own parameter set so multiple tiers of reference have their own utility. While these facets of the meta-database do not necessarily detract from the model comparisons or other results in this work, they should always be kept in mind when conducting any sort of validation or characterization of a potential model.

The emphasis of this research was on the predictive behavior of the longitudinal ML models themselves, against each other and the inferential reference designs. One aspect not addressed was feature selection or identification of the most powerful predictive covariates which is a common emphasis in machine learning research. This has practical implications for ADRD applications specifically when considering the potential role of subject demographic characteristics which have less prevalent coverage. For example, genotyping of the *APOE4* allele was by far the most restrictive covariate of the meta-database and contributed to the greatest reduction of subjects for analysis. However, this allele is well-known for its association with idiopathic AD and could very easily be one of the most powerful predictors of cognitive outcomes. Although genotyping *APOE4* is not especially common in clinical settings, providers may be more inclined to determine allele counts if its utility in predicting impairment were explicitly demonstrated, especially when forecasting future observations. But as mentioned, covariate identification was not this study's goal which focused on model characterization over feature selection. This current characterization of the models themselves is a critical first step in the assessment process which gives results such as identifying the MERF model as one of the most powerful predictors of ADAS-Cog. The impact of feature space is, in turn, a natural follow-up to the current work, especially in light of the results of aim 3 which indicated similar prediction errors and biases between the inferential BR and ML MERF models when given the same restricted feature space. Wider covariates sets, with an emphasis on meaningful identification, would provide more detail information on which contexts and setting are best suited for ML methods in ADRD research.

A similar expansion is less focused on ML methods per se, but rather the role of cut point optimization for classification tasks in general. As discussed, many of the evaluation metrics for classification are reliant on translating a score value to a binary outcome which confounds observations of reduced recall and high precision. This is especially true in the case of imbalanced datasets such as the meta-database. Improvements in AUC of the ROC curve, which is agnostic to strict classification decisions, offset this concern to a degree but use of different cut point optimization techniques and the inclusion of other evaluation metrics which are less reliant on confusion matrix outcomes could be greatly beneficial. However, that is not to say translation from a response score to a strict categorization should be abandoned altogether. From a clinical perspective, classification to a label such as impaired or unimpaired is essentially the ultimate goal for these tasks, especially at ambiguous threshold scores, and should be a focus of ML classification methods research. Accordingly, cut point optimizations and the impact they have on the robustness of machine learning models is also a worthwhile pursuit.

One of the most unexpected and intriguing results was the comparison of the whole subject trajectory predictions for the non-sequential FNN models when up against the 1D CNN and LSTM RNN models in aim 2, especially for the ADAS-Cog. As mentioned, these results could be cast one of two ways: either why did the FNN models perform as well as they did or, conversely, why did the sequential neural networks perform relatively poorly in comparison. It is hard to claim neural networks are ill-suited for whole subject trajectory prediction given the results of the FNN. In addition, the utility of these specific sequential adaptations must be acknowledged as they were some of the most powerful methods for final observation forecasting. A potential rationale is not with the models themselves but rather how the trajectory predictions were calculated for the CNN and LSTM designs. While the use of a baseline FNN model to seed the trajectory predictions was certainly a sufficient and feasible solution to the empty array issue, there may be alternative methods that provide more accurate predictions and would be less prone to propagation of errors due to the iterative response calculations of the sequential ANN methods. Although the predictive capacity of the sequential models cannot be denied when using known sequence data for forecasting future observations, investigation into alternative prediction methods for trajectories is a natural follow-up in neural network applications to cognitive outcomes.

A final point is the role of performance evaluation and hyperparameter selection for machine learning. Hyperparameters are a unique aspect in ML methods as they cannot be directly determined from the data. Model tuning and hyperparameter selection was certainly an aspect of this study and helped provide even better predictions of the cognitive outcomes, but it was not a central component. Expansion of the grid searches with additional hyperparameters using more settings could easily lead to further refinements and result in even better performance of the models. In addition, while some hyperparameters assist with predictive ability of the models, as was their utilization goal here, others instead focus on improving the performance of the models by reducing processing time or requisite memory. These performance aspects were not explicitly evaluated during this study but are well worth characterization as well. The most obvious example is the boosted mixed-effects trees which although demonstrating high predictive performance, including the highest ROC AUC when predicting whole trajectories of CDR impairment, unquestionably presented as the slowest and most resource intensive method of this study. As mentioned, it was so demanding it was the only model to use 5-fold cross-validation while all other models used 10-fold. Although its ROC AUC on CDR impairment forecasting was demonstrated as significantly superior to most models it cannot be reasonably recommended as a method with the current implementation. Other models still perform well but at a fraction of the cost of time and resources. Performance outcomes are just as critical as predictive ability with machine learning models and fully evaluating an individual model on both aspects is key. The translation of any method to a practical setting should always be the goal in this kind of methods research and is important to remember when applying these implementations to clinical ADRD research.

CONCLUSIONS

One of the central themes of this study was to bridge the divide between statistical consideration and clinical utility of longitudinal machine learning in Alzheimer's disease research by providing an overview of how these methods can be applied and what investigators can expect in terms of performance. In addition to statistical considerations of these models, a goal was to demonstrate the specific practical utility of the methods and how they can benefit AD researchers and clinicians in their own studies. The most important caveat is there was no general consensus in the best model for either regression or classification in these designs and each of these models has their own strengths and weaknesses. Whether being used for whole subject trajectories or forecasting future data points within an individual, the preferred model requires an informed decision and depends on the goals of the investigator or clinician. This also relates to goals of what aspect of prediction is being optimized such as minimization of prediction bias even at the cost of additional error for ADAS-Cog or protection specifically against false negatives when classifying impairment from the CDR. This further extends into how subject-specific effects should be utilized and the role of imputation can play, including its efficacy in developing robust synthetic cohorts for feasibility but generally detracting from prediction on a subject-by-subject basis. If the goal is generating data where on average accuracy is desired for a full cohort, the inclusion of imputed subject-specific effects is warranted. However, if the goal is instead to predict the trajectory or endpoint of a specific individual, then only population level fixed effects should be used to get the most accurate prediction possible. Recognizing these scenarios and aligning them with the desired study design and clinical goals is what will lead to better, more accurate prediction of outcomes and higher quality research in cognitive decline. Ultimately, this dissertation is a critical first step in characterizing predictive performance of longitudinal machine learning methods in Alzheimer's disease, serving as both a current evaluation of the field and a pipeline for evaluation for future statistical learning paradigms.

REFERENCES

Ackley, D., Hinton, G. E., & Sejnowski, T. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science, 9*(1), 147-169.

Allaire, J., & Chollet, F. (2021). *keras: R Interface to 'Keras'. R package version 2.4.0.* https://CRAN.R-project.org/package=keras

Allaire, J., & Yuan, T. (2021). *tensorflow: R Interface to 'TensorFlow'. R package version 2.4.0.* https://CRAN.R-project.org/package=tensorflow

Ard, M. C., Raghavan, N., & Edland, S. D. (2015, Sep-Oct). Optimal composite scores for longitudinal clinical trials under the linear mixed effects model. *Pharm Stat, 14*(5), 418-426. https://doi.org/10.1002/pst.1701

Bader, J. M., Geyer, P. E., Muller, J. B., Strauss, M. T., Koch, M., Leypoldt, F., Koertvelyessy, P., Bittner, D., Schipke, C. G., Incesoy, E. I., Peters, O., Deigendesch, N., Simons, M., Jensen, M. K., Zetterberg, H., & Mann, M. (2020, Jun). Proteome profiling in cerebrospinal fluid reveals novel biomarkers of Alzheimer's disease. *Mol Syst Biol, 16*(6), e9356. https://doi.org/10.15252/msb.20199356

Barber, R. F., Reimherr, M., & Schill, T. (2017). The function-on-scalar LASSO with applications to longitudinal GWAS. *Electronic Journal of Statistics, 11*(1), 1351-1389.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *J Stat Software, 67*(1), 1-48.

Bernal-Rusiel, J. L., Greve, D. N., Reuter, M., Fischl, B., Sabuncu, M. R., & Alzheimer's Disease Neuroimaging, I. (2013, Feb 1). Statistical analysis of longitudinal neuroimage data with Linear Mixed Effects models. *Neuroimage, 66*, 249-260. https://doi.org/10.1016/j.neuroimage.2012.10.065

Bhatnagar, S., Oualkacha, K., Yang, Y., & Greenwood, C. (2021). *ggmix: Variable Selection in Linear Mixed Models for SNP Data. R package version 0.0.2*. https://CRAN.R-project.org/package=ggmix

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer Publishing.

Breiman, L. (1996). Bagging Predictors. *Machine Learning, 24*(2), 123-140.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5-32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Wadsworth & Brooks / Cole Advanced Books.

Breitve, M. H., Chwiszczuk, L. J., Bronnick, K., Hynninen, M. J., Auestad, B. H., Aarsland, D., & Rongve, A. (2018). A Longitudinal Study of Neurocognition in Dementia with Lewy Bodies Compared to Alzheimer's Disease. *Front Neurol, 9*, 124. https://doi.org/10.3389/fneur.2018.00124

Burton, P., Gurrin, L., & Sly, P. (1998, Jun 15). Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Stat Med, 17*(11), 1261-1291. https://doi.org/10.1002/(sici)1097-0258(19980615)17:11

Capitaine, L., Genuer, R., & Thiebaut, R. (2021, Jan). Random forests for high-dimensional longitudinal data. *Stat Methods Med Res, 30*(1), 166-184. https://doi.org/10.1177/0962280220946080

Capuano, A. W., Wilson, R. S., Leurgans, S. E., Dawson, J. D., Bennett, D. A., & Hedeker, D. (2018, Mar). Sigmoidal mixed models for longitudinal data. *Stat Methods Med Res, 27*(3), 863-875. https://doi.org/10.1177/0962280216645632

Chen, S., & Bowman, F. D. (2011, Dec). A Novel Support Vector Classifier for Longitudinal High-dimensional Data and Its Application to Neuroimaging Data. *Stat Anal Data Min, 4*(6), 604-611. https://doi.org/10.1002/sam.10141

Chen, S., Grant, E., Wu, T. T., & Bowman, F. D. (2014, Jan). Statistical Learning Methods for Longitudinal High-dimensional Data. *Wiley Interdiscip Rev Comput Stat, 6*(1), 10-18. https://doi.org/10.1002/wics.1282

Chen, T., Zeng, D., & Wang, Y. (2015, Dec). Multiple kernel learning with random effects for predicting longitudinal outcomes and data integration. *Biometrics, 71*(4), 918-928. https://doi.org/10.1111/biom.12343

Chen, Y. F., Ni, X., Fleisher, A. S., Zhou, W., Aisen, P., & Mohs, R. (2018). A simulation study comparing slope model with mixed-model repeated measure to assess cognitive data in clinical trials of Alzheimer's disease. *Alzheimers Dement (N Y), 4*, 46-53. https://doi.org/10.1016/j.trci.2017.12.002

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297.

De Velasco Oriol, J., Vallejo, E. E., Estrada, K., Tamez Pena, J. G., & Disease Neuroimaging Initiative, T. A. (2019, Dec 16). Benchmarking machine learning models for late-onset alzheimer's disease prediction from genomic data. *BMC Bioinformatics, 20*(1), 709. https://doi.org/10.1186/s12859-019-3158-x

Di, J., Wang, D., Brashear, H. R., Dragalin, V., & Krams, M. (2016, Mar-Apr). Continuous event monitoring via a Bayesian predictive approach. *Pharm Stat, 15*(2), 109-122. https://doi.org/10.1002/pst.1727

Donohue, M. C., & Aisen, P. S. (2012, Apr). Mixed model of repeated measures versus slope models in Alzheimer's disease clinical trials. *J Nutr Health Aging, 16*(4), 360-364. https://doi.org/10.1007/s12603-012-0047-7

Doody, R. S., Massman, P., & Dunn, J. K. (2001, Mar). A method for estimating progression rates in Alzheimer disease. *Arch Neurol, 58*(3), 449-454. https://doi.org/10.1001/archneur.58.3.449

Du, W., Cheung, H., Johnson, C. A., Goldberg, I., Thambisetty, M., & Becker, K. (2015). A longitudinal support vector regression for prediction of ALS score. 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),

Eliot, M., Ferguson, J., Reilly, M. P., & Foulkes, A. S. (2011). Ridge regression for longitudinal biomarker data. *The International Journal of Biostatistics, 7*(1).

Fisher, C. K., Smith, A. M., Walsh, J. R., & Coalition Against Major, D. (2019, Sep 20). Machine learning for comprehensive forecasting of Alzheimer's Disease progression. *Sci Rep, 9*(1), 13622. https://doi.org/10.1038/s41598-019-49656-2

Fitzmaurice, G., Laird, N., & Ware, J. (2011). *Applied Longitudinal Analysis, 2nd Edition*. John Wiley & Sons.

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018, Oct). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behav Res Methods, 50*(5), 2016-2034. https://doi.org/10.3758/s13428-017-0971-x

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. Proc 13th Int Conf on Machine Learning,

Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machines. *Annal Stat, 29*, 1189-1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Software, 33*(1), 1-22.

Fuegi, J., & Francis, J. (2003). Lovelace & Babbage and the creation of the 1843 'notes'. *IEEE Annals of the History of Computing, 25*(4), 16-26.

Fukushima, K. (1980). A self-organizing neural network model for a mechansim of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*(4), 193-202.

Gavidia-Bovadilla, G., Kanaan-Izquierdo, S., Mataro-Serrat, M., Perera-Lluna, A., & Alzheimer's Disease Neuroimaging, I. (2017). Early Prediction of Alzheimer's Disease Using Null Longitudinal Model-Based Classifiers. *PLoS One, 12*(1), e0168011. https://doi.org/10.1371/journal.pone.0168011

Giil, L. M., & Aarsland, D. (2020). Greater Variability in Cognitive Decline in Lewy Body Dementia Compared to Alzheimer's Disease. *J Alzheimers Dis, 73*(4), 1321-1330. https://doi.org/10.3233/JAD-190731

Giil, L. M., Aarsland, D., & Vik-Mo, A. O. (2021). Differentiating traits and states identifies the importance of chronic neuropsychiatric symptoms for cognitive prognosis in mild dementia. *Alzheimers Dement (Amst), 13*(1), e12152. https://doi.org/10.1002/dad2.12152

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep Sparse Rectifier Neural Networks. *PMLR, 15*, 315-323.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Groll, A. (2017). *glmmLasso: Variable Selection for Generalized Linear Mixed Models by L1-Penalized Estimation. R package version 1.5.1.* https://CRAN.R-project.org/package=glmmLasso

Guo, J., Shang, Y., Fratiglioni, L., Johnell, K., Welmer, A. K., Marseglia, A., & Xu, W. (2021, Mar 26). Individual changes in anthropometric measures after age 60 years: a 15-year longitudinal population-based study. *Age Ageing*. https://doi.org/10.1093/ageing/afab045

Hajjem, A., Bellavance, F., & Larocque, D. (2011). Mixed effects regression trees for clustered data. *Statistics & probability letters, 81*(4), 451-459.

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation, 84*(6), 1313-1328.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning, 2nd Ed*. Springer.

Hickey, J., Metcalfe, P., Ridgeway, G., Schroedl, S., Southworth, H., & Therneau, T. (2016). *gbm: Genealized Boosted Regression Models*. https://github.com/gbm-developers/gbm

Higgins, C. (2017). *A Brief History of Deep Blue, IBM's Chess Computer*. https://www.mentalfloss.com/article/503178/brief-history-deep-blue-ibms-chess-computer

Higgins, J. P., Whitehead, A., Turner, R. M., Omar, R. Z., & Thompson, S. G. (2001, Aug 15). Meta-analysis of continuous outcome data from individual patients. *Stat Med, 20*(15), 2219-2241. https://doi.org/10.1002/sim.918

Hochreiter, S. (1991). *Untersuchungen zu dynamischen neuronalen Netzen (German)* Technische Univ. Munich]. Munich, Germany.

Hochreiter, S. (1998). The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, 6*(2), 107-116.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735-1780.

Hoerl, A., & Kennard, R. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics, 12*(1), 55-67.

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *PNAS, 79*(8), 2554-2558.

Huang, X., Liu, H., Li, X., Guan, L., Li, J., Tellier, L., Yang, H., Wang, J., & Zhang, J. (2018, Jan 10). Revealing Alzheimer's disease genes spectrum in the whole-genome by machine learning. *BMC Neurol, 18*(1), 5. https://doi.org/10.1186/s12883-017-1010-3

Ito, K., Corrigan, B., Romero, K., Anziano, R., Neville, J., Stephenson, D., & Lalonde, R. (2013). Understanding placebo responses in Alzheimer's disease clinical trials from the literature meta-data and CAMD database. *J Alzheimers Dis, 37*(1), 173-183. https://doi.org/10.3233/JAD-130575

Ivakhnenko, A. (1968). The Group Method of Data Handling - a Rival of the Method of Stochastic Apprxoimation. *Soviet Automatic Control, 13*(3), 43-55.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the Limits of Language Modeling. *arXiv 1602.02410*.

Karch, J. D., Brandmaier, A. M., & Voelkle, M. C. (2020). Gaussian Process Panel Modeling-Machine Learning Inspired Analysis of Longitudinal Panel Data. *Front Psychol, 11*, 351. https://doi.org/10.3389/fpsyg.2020.00351

Kaur, H., Singh, Y., Singh, S., & Singh, R. B. (2021, Apr). Gut microbiome-mediated epigenetic regulation of brain disorder and application of machine learning for multi-omics data analysis. *Genome, 64*(4), 355-371. https://doi.org/10.1139/gen-2020-0136

Kelley, H. (1960). Gradient Theory of Optimal Fight Paths. *American Rocket Society Journal, 30*(10), 947-954.

Kennedy, R. E., Cutter, G. R., & Schneider, L. S. (2014, May). Effect of APOE genotype status on targeted clinical trials outcomes and efficiency in dementia and mild cognitive impairment resulting from Alzheimer's disease. *Alzheimers Dement, 10*(3), 349-359. https://doi.org/10.1016/j.jalz.2013.03.003

Kim, S. E., Lee, B., Jang, H., Chin, J., Khoo, C. S., Choe, Y. S., Kim, J. S., Kang, S. H., Kim, H. R., Hwangbo, S., Jeong, J. H., Yoon, S. J., Park, K. W., Kim, E. J., Yoon, B., Jang, J. W., Hong, J. Y., Na, D. L., Seo, S. W., Choi, S. H., & Kim, H. J. (2021, Feb 19). Cognitive trajectories of patients with focal ss-amyloid deposition. *Alzheimers Res Ther, 13*(1), 48. https://doi.org/10.1186/s13195-021-00787-7

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., & Inman, D. (2021). 1D convolutional neural networks and applications: A survey. *Mechanical Systems and SIgnal Processing, 151*, 107398.

Klen, R., Karhunen, M., & Elo, L. L. (2020, Jan 23). Likelihood contrasts: a machine learning algorithm for binary classification of longitudinal data. *Sci Rep, 10*(1), 1016. https://doi.org/10.1038/s41598-020-57924-9

Klöppel, S., Stonnington, C. M., Barnes, J., Chen, F., Chu, C., Good, C. D., Mader, I., Mitchell, L. A., Patel, A. C., Roberts, C. C., Fox, N. C., Jr, C. R. J., Ashburner, J., & Frackowiak, R. S. J. (2008). Accuracy of dementia diagnosis: a direct comparison between radiologists and a computerized method. *Brain, 131*(11), 2969-2974.

Lang, A., Carass, A., Al-Louzi, O., Bhargava, P., Solomon, S. D., Calabresi, P. A., & Prince, J. L. (2016). Combined registration and motion correction of longitudinal retinal OCT data. Medical Imaging 2016: Image Processing,

Le, Q. V., Jaitly, N., & Hinton, G. E. (2015). A simple way to initialize recurrent networks of rectified linear units. *arXiv preprint arXiv:1504.00941*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015, May 28). Deep learning. *Nature, 521*(7553), 436-444. https://doi.org/10.1038/nature14539

Lee, S., Yoon, S., & Cho, H. (2017). Human activity recognition from accelerometer data using Convolutional Neural Network. 2017 IEEE International Conference on Big Data and Smart Computing,

Leondes, C. (2001). *Expert Systems: The Technology of Knowledge Management and Decision Making for the 21st Century*. Academic Press.

Li, Q., Guo, Y., He, Z., Zhang, H., George, T. J., Jr., & Bian, J. (2020). Using Real-World Data to Rationalize Clinical Trials Eligibility Criteria Design: A Case Study of Alzheimer's Disease Trials. *AMIA Annu Symp Proc, 2020*, 717-726. https://www.ncbi.nlm.nih.gov/pubmed/33936446

Liu, C.-L., Hsaio, W.-H., & Tu, Y.-C. (2018). Time series classification with multivariate convolutional neural network. *IEEE Transactions on Industrial Electronics, 66*(6), 4788-4797.

Liu, X., Zhang, F., Hou, Z., Wang, Z., Mian, L., Zhang, J., & Tang, J. (2020). Self-supervised Learning: Generative or Contrastive. *arXiv, 2006.08218*, 1-20.

Luts, J., Molenberghs, G., Verbeke, G., Van Huffel, S., & Suykens, J. A. (2012). A mixed effects least squares support vector machine model for classification of longitudinal data. *Computational Statistics & Data Analysis, 56*(3), 611-628.

Marti-Juan, G., Sanroma-Guell, G., & Piella, G. (2020, Jun). A survey on machine and statistical learning for longitudinal analysis of neuroimaging data in Alzheimer's disease. *Comput Methods Programs Biomed, 189*, 105348. https://doi.org/10.1016/j.cmpb.2020.105348

McCulloch, W., & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics, 5*, 115-133.

Miller, P., & McArtor, D. (2017). *mvtboost: Tree Boosting for Multivariate Outcomes*. https://github.com/patr1ckm/mvtboost

Miller, P. J., McArtor, D. B., & Lubke, G. H. (2017). metboost: Exploratory regression analysis with hierarchically clustered data. *arXiv preprint arXiv:1702.03994*.

Milliken, J. K., & Edland, S. D. (2000, Jun 15-30). Mixed effect models of longitudinal Alzheimer's disease data: a cautionary note. *Stat Med, 19*(11-12), 1617-1629. https://doi.org/10.1002/(sici)1097-0258(20000615/30)19:11/12<1617::aid-sim450>3.0.co;2-c

Minsky, M., & Papert, S. (1969). *Perceptrons: an introduction to computational geometry*. MIT Press.

Mirzaei, G., Adeli, A., & Adeli, H. (2016, Dec 1). Imaging and machine learning techniques for diagnosis of Alzheimer's disease. *Rev Neurosci, 27*(8), 857-870. https://doi.org/10.1515/revneuro-2016-0029

Naik, B., Mehta, A., & Shah, M. (2020, Nov 5). Denouements of machine learning and multimodal diagnostic classification of Alzheimer's disease. *Vis Comput Ind Biomed Art, 3*(1), 26. https://doi.org/10.1186/s42492-020-00062-w

Ngufor, C., Van Houten, H., Caffo, B. S., Shah, N. D., & McCoy, R. G. (2019, Jan). Mixed effect machine learning: A framework for predicting longitudinal change in hemoglobin A1c. *J Biomed Inform, 89*, 56-67. https://doi.org/10.1016/j.jbi.2018.09.001

Nowok, B., Raab, G., & Dibben, C. (2016). synthpop: Bespoke Creation of Synthetic Data in R. *J Stat Software, 74*(11), 1-26.

O'Shea, D. M., Thomas, K. R., Asken, B., Lee, A. K. W., Davis, J. D., Malloy, P. F., Salloway, S. P., Correia, S., & Alzheimer's Disease Neuroimaging, I. (2021). Adding cognition to AT(N) models improves prediction of cognitive and functional decline. *Alzheimers Dement (Amst), 13*(1), e12174. https://doi.org/10.1002/dad2.12174

Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X., & Ward, R. (2016). Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval. *arXiv, 1502.06922v3*, 1-25.

Pande, A., Li, L., Rajeswaran, J., Ehrlinger, J., Kogalur, U. B., Blackstone, E. H., & Ishwaran, H. (2017). Boosted multivariate trees for longitudinal data. *Machine Learning, 106*(2), 277-305.

Pedregosa, F., Varoquaux, G., Gramfort, A., MMichel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. *JMLR, 12*, 2825-2830.

Pepe, M. S., Fan, J., Feng, Z., Gerds, T., & Hilden, J. (2015, Oct 1). The Net Reclassification Index (NRI): a Misleading Measure of Prediction Improvement Even with Independent Test Data Sets. *Stat Biosci, 7*(2), 282-295. https://doi.org/10.1007/s12561-014-9118-0

Polhamus, D. (2013). adsim: Simulate Alzhieimer's Disease clincial trials. R package version 3.0.

Raina, R., Madhavan, A., & Ng, A. (2009). Large-scale Deep Unsupervised Learning using Graphics Processors. Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada.

Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A., & Davatzikos, C. (2017, Jul 15). A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *Neuroimage, 155*, 530-548. https://doi.org/10.1016/j.neuroimage.2017.03.057

Ravi, D., Wong, C., Lo, B., & Yang, G.-Z. (2016). A deep learning approach to on-node sensor data analytics for mobile or wearable devices. *IEEE journal of biomedical and health informatics, 21*(1), 56-64.

Rogers, J. A., Polhamus, D., Gillespie, W. R., Ito, K., Romero, K., Qiu, R., Stephenson, D., Gastonguay, M. R., & Corrigan, B. (2012, Oct). Combining patient-level and summary-level data for Alzheimer's disease modeling and simulation: a beta regression meta-analysis. *J Pharmacokinet Pharmacodyn, 39*(5), 479-498. https://doi.org/10.1007/s10928-012-9263-3

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review, 65*(6), 386-408.

Russel, S., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach (2nd ed)*. Prentice Hall.

Schelldorfer, J. (2011). *lmmlasso: Linear mixed-effects models with Lasso. R package version 0.1-2.* https://CRAN.R-project.org/package=lmmlasso

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks, 61*, 85-117.

Schubert, B., & Marks, D. (2017). *GELMMnet - Generalized network-based elastic-net linear mixed model*. https://github.com/debbiemarkslab/GELMMnet

Sidi, J. (2020). *lmmen: Linear Mixed Model Elastic Net. R package version 1.0.* https://CRAN.R-project.org/package=lmmen

Singstad, B., & Tronstad, C. (2020). Convolutional Neural Network and Rule-Based Algorithms for Classifying12-lead ECGs. *Computing in Cardiology, 47*, 1-4.

Skolov, A., Carlin, D. E., Paull, E. O., Baertsch, R., & Stuart, J. M. (2016). Pathway-based genomics prediction using generalized elastic net. *PLoS Computational Biology, 12*(3), e1004790.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the lasso. *J of the Royal Statistical Society, 58*(1), 267-288.

Ushey, K., Allaire, J., & Tang, Y. (2021). *reticulate: Interace to 'Python'. R package version 1.20*. https://CRAN.R-project.org/package=reticulate

Uspenskaya-Cadoz, O., Alamuri, C., Wang, L., Yang, M., Khinda, S., Nigmatullina, Y., Cao, T., Kayal, N., O'Keefe, M., & Rubel, C. (2019). Machine Learning Algorithm Helps Identify Non-Diagnosed Prodromal Alzheimer's Disease Patients in the General Population. *J Prev Alzheimers Dis, 6*(3), 185-191. https://doi.org/10.14283/jpad.2019.10

Vapnik, V., & Chervonenkis, A. (1974). *Pattern Recognition Theory, Statistical Learning Problems*. Nauka.

Wang, G., Berry, S., Xiong, C., Hassenstab, J., Quintana, M., McDade, E. M., Delmar, P., Vestrucci, M., Sethuraman, G., Bateman, R. J., & Dominantly Inherited Alzheimer Network Trials, U. (2018, Sep 20). A novel cognitive disease progression model for clinical trials in autosomal-dominant Alzheimer's disease. *Stat Med, 37*(21), 3047-3055. https://doi.org/10.1002/sim.7811

Wilson, C., & Li, K. (2019). *RMKL: Multiple Kernel Learning for Classification or Regression Problems. R package version 1.0.* https://CRAN.R-project.org/package=RMKL

Xu, G., Ren, T., Chen, Y., & Che, W. (2020). A One-Dimensional CNN-LSTM Model for Epileptic Seizure Recognition Using EEG Signal Analysis. *Fronteirs in Neuroscience, 14*, 1-9.

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *J of the Royal Statistical Society, 67*(2), 301-320.

APPENDIX A

GLOSSARY OF TERMS

Alzheimer's disease

A type of progressive dementia affecting memory, thinking and behavior which eventually interferes with daily tasks. Pathologically characterized by the presence of β-amyloid plaques, neurofibrillary tangles of phosphorylated tau, and neurodegenerative loss of brain tissue. Clinical dementia is largely defined by the presence of functional impairment in addition to cognitive impairment.

Alzheimer's Disease Assessment Scale – Cognitive Subscale

One half of the Alzheimer's Disease Assessment Scale assessing the severity of the cognitive symptoms of dementia. Comprised of 11 tasks for both subject-completion and observer-based assessment evaluating memory, language, and praxis. Generally used in clinical trials or other research environments and not in clinical practice.

Alzheimer's Disease Neuroimaging Initiative

A long-standing collaborative study initiated in 2004 to study Alzheimer's disease by identifying more sensitive and accurate biomarkers, most notably imaging measures like magnetic resonance imaging and positron emission tomography. This observational study has undergone several phases and extensions with the most recent having begun in 2016.

Alzheimer's disease and related dementias

A collection of dementia diagnoses sharing many of the cognitive and pathological features with Alzheimer's disease which makes distinguishing the different diseases difficult. Includes disorders such as frontotemporal degeneration, Lewy body dementia, vascular contributions to cognitive impairment and dementia, mixed etiology dementias and others.

apolipoprotein E

A protein involved in the metabolism of fats. Has three major alleles within humans (E2, E3, and E4) with the E4 variant highly associated with idiopathic Alzheimer's disease with both increased prevalence and an earlier age of onset. Dosage effects are also observed with more pronounced effects seen by homozygous carriers of the E4 subtype compared to heterozygous carriers.

artificial neural network

See **neural network**.

bagging

A portmanteau of bootstrap and aggregation. An ensemble method which samples from the training data before tree building. This is repeated several times with the trees' functions averaged together at the end e.g. majority classification vote or mean response.

bias-variance trade-off

A property of predictive models where total prediction error is a function of both bias and variance. A high bias model is unable to sufficiently model the relationship between training and testing data (underfitting) while a high variance model erroneously models noise in the training data and is unable to generalize to other data (overfitting).

boosting

In this context, training data is fit to a single classifier (a weak learner) which modifies the data e.g. calculation and application of model residuals. This modified data is then fit to a new tree and the process is sequentially iterated to eventually lead to a strong learner.

bootstrap

The technique of randomly sampling from a dataset with replacement. Can also refer to the resulting dataset. Creates a dataset with an empirical distribution expected to be similar to the original data's probability distribution.

Clinical Dementia Rating

An instrument used in both clinical practice and research for staging of Alzheimer's disease, assessing both cognitive ability as well as basic functions of daily living and engagement. Comprises six domains (memory, orientation, judgment, community affairs, home & hobbies, personal care) with each having individual scores as well as a global score of 0, 0.5, 1, 2, and 3. The global score in turn corresponds with an Alzheimer's disease stage of normal/intact, mild cognitive impairment, and mild/moderate/severe dementia.

cognitively intact

Patients with no evidence of cognitive decline beyond what would be expected in their age group. Pathological changes are generally not considered when defining normal cognition although further subsets are possible. Also referred to as cognitively normal.

convolutional neural network

A neural network where connections between layers are limited to local sets of incoming nodes, allowing for fewer nodes at each layer instead of the fully connected layers seen in most neural networks. This limits connections to emphasize local features within a "window" with less attention paid to more distal features. In the one-dimensional case can be used for longitudinal data by focusing on proximal events.

Critical Paths for Alzheimer's Disease

A consortium project created to develop new tools and methods to assist in the design of clinical interventional trials in Alzheimer's disease with an emphasis on drug interventions. One such tool is a simulation framework of drug-based clinical trials using the Alzheimer's Disease Assessment Scale – Cognitive Subscale as an outcome measure.

cross-validation

A type of evaluation where a training set is split into a pre-specified number of folds. Separate models are developed with each fold excluded in sequence with the holdout then used for evaluation purposes for things like hyperparameter optimization. Final evaluations are often done across the set of developed models with results then aggregated together.

decision tree

Any sort of flow-like structure used to support decision making. Comprised of nodes and edges where a node consists of some type of function and the edge is the decision made on the result of that function leading to subsequent nodes or a final decision (leaf node). Also qualifies as a type of machine learning algorithm.

discriminative model

Models which learn classification boundaries or response values using conditional probability of the target Y given the input X. Is generally associated with supervised learning.

elastic net

Regularization which uses a linear combination of $\ell1$ and $\ell2$ loss. Is designed to overcome limitations of other regularization methods such as feature selection issues.

ensemble

A machine learning method which uses multiple algorithm implementations in combination or aggregation to improve prediction performance beyond that of any single component algorithm or model.

features

The collection of input variables used for training a machine learning or inferential regression model.

feed-forward neural network

A rudimentary type of neural network which follows a very basic design of nodes fully connected across one or more layers without further modification or accommodation to various data structures.

generalized linear mixed model tree

A type of decision tree which accounts for longitudinal or panel data by applying a linear mixed model in the terminal leaf nodes.

generative model

Models which learn boundaries or response estimates by directly modelling the joint distribution of X and Y before applying Bayes's rule. In the current context generally refers to unsupervised or self-supervised methods.

hyperparameters

A specific type of parameter that is used to control the learning process in order to improve model performance. Unlike trained parameters, these are estimated during validation and are generally not defined by data themselves.

kernel method

A method of determining a hyperplane used for maximal margin modelling which maps the features to a higher dimensional space to achieve separation. Specifically uses kernel functions to implicitly calculate the higher dimensional feature space using the inner products of data point pairs.

least absolute shrinkage and selection operator

A regularization method which uses a variation on the sum of the absolute value of parameter estimates as the loss function called the $\ell 1$ loss. Uses a shrinkage term as a hyperparameter to control the penalization.

linear mixed model

Also referred to as linear mixed-effects models. See **mixed-effects regression**.

long-short term memory

A variation of recurrent neural networks developed to address some limitations with training and parameter estimation. Specifically allows for layers to retain or forget previous data in a variable fashion and incorporate new inputs stochastically to update activation functions accordingly.

longitudinal

Defined here as any type of data or method which involves repeated measurements of the same variable on the same unit under a time series framework e.g. annual measures on a patient. Also referred to as panel data or panel method.

mild cognitive impairment

Generally seen as a transitional state between natural cognitive decline in aging and the earliest features of dementia. Sometimes referred to as prodromal Alzheimer's disease. Memory deficits are present but do not yet interfere with daily living.

machine learning

The utilization of algorithms to allow systems to iteratively and automatically improve models through exposure and experience. Often contrasts with traditional inferential statistics due to greater emphasis on prediction rather than hypothesis testing and covariate interpretation.

Mini-Mental State Exam

A 30-point test which measures cognitive ability with lower scores corresponding to greater levels of impairment. Can also be used as a staging instrument although consensus of diagnosis to level of disease is less well-defined compared to the Clinical Dementia Rating as it does not address impairment of function beyond cognitive ability. In general, scores of 28 or higher are accepted as cognitively intact, 23-28 indicate mild cognitive impairment, and scores below 23-21 are associated with varying levels of dementia.

mixed-effects regression

A model comprised of both fixed (i.e. measured) effects and random (i.e. unobserved) effects. Commonly used in the analysis of longitudinal or panel data.

multi-layer perceptron

See **feed-forward neural network**.

neural network

Defined here as any machine learning method which feeds input data through a set of layers, each comprised of several nodes, before the final output decision. Each node involves some sort of activation function with weights connecting to the nodes of subsequent layers determining how data is passed through the layers and evaluated.

panel

See **longitudinal**.

parameters

The collection of model-specific function values estimated during training which the machine learning model uses for response prediction.

penalization

See **regularization**.

radial basis function

A kernel in the form of a radial or Gaussian function to allow a mapping to an infinite dimensional space. One of the more common kernels used in support vector machines.

random effects

A model parameter which is itself a random variable. Often used in hierarchical, panel, or clustered designs. For linear mixed models, often used to model unit-specific longitudinal variations.

random forest

A variation on bagging which instead of sampling from the subjects or units will sample the number of features used to build any given tree. These random trees are then aggregated but have the benefit of being uncorrelated.

recurrent neural network

A type of neural network which allows previous layer outputs to be used as inputs for all subsequent layers. Allows for having a variable number of layers which does not require a fixed size for the input and leverages series structure by using prior data.

regularization

A class of technique used in regression modelling to discourage a more complex model by constraining the estimates of a model parameter. The general goal is to further minimize total prediction error by greatly reducing variance at the cost of slightly higher bias in the model. Also referred to as penalization.

ridge regression

A method which uses a variation on the sum of the squared parameter estimates as the loss function called the $\ell 2$ loss. Regularizes a model using a shrinkage hyperparameter but will not allow parameter estimates to completely go to zero such that all features remain within the model.

statistical learning

See **machine learning**.

self-supervised learning

A mixture of supervised and unsupervised learning where only a small portion of the training data has known labels. The goal is a mixture of supervised and unsupervised techniques such as pre-training a network with unlabeled data for initialization and then estimating and fine-tuning the model parameters with the labeled input data.

supervised learning

A machine learning method where the responses of the testing set are known prior to training, allowing direct mapping of the input variables to the response variables. Thus, the goal is to estimate the mapping function as to be applied to new input data.

support vector machine

A machine learning technique which uses a maximal margin classifier to determine a response (support vector classifier) by separating data using a hyperplane. This is combined with some type of non-linearity applied to the hyperplane to create the model.

testing set

A completely held-out dataset used for unbiased evaluation of the final model after both training and tuning.

training set

The dataset used to fit the feature parameters of a supervised or semi-supervised machine learning system using an algorithm's optimization method.

validation set

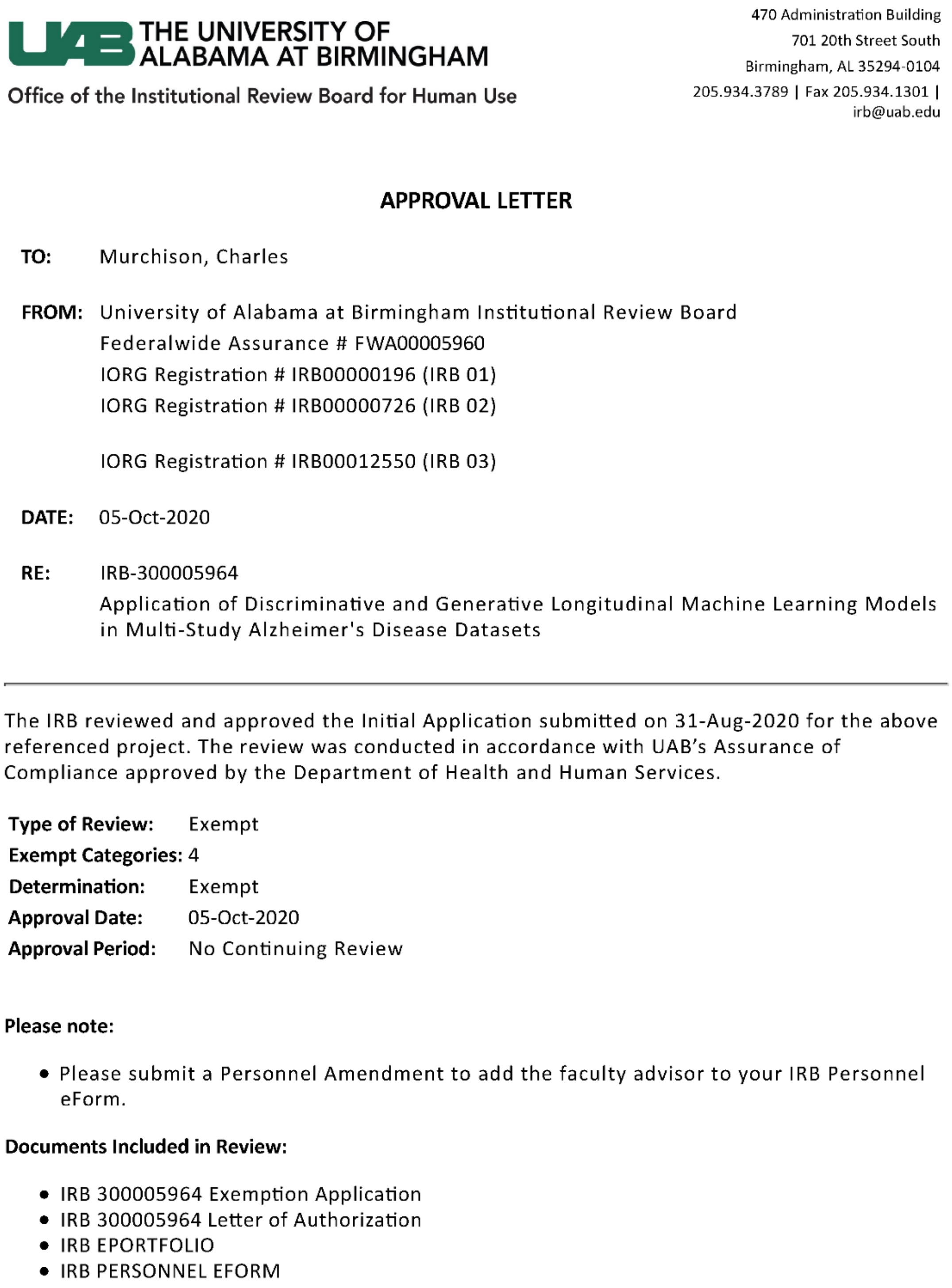
A dataset held out during the training portion which is used to optimize or tune hyperparameters and provide an unbiased evaluation of a machine learning model's performance. Often taken as a subset of a training set using methods such as cross-validation.

unsupervised learning

A machine learning method where the labels or values of the response variables are unknown. Instead of approximating the mapping the labelling function directly the goal is to model the underlying structure of the input data.

APPENDIX B

IRB APPROVAL LETTER



THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Office of the Institutional Review Board for Human Use

470 Administration Building
701 20th Street South
Birmingham, AL 35294-0104
205.934.3789 | Fax 205.934.1301 | irb@uab.edu

APPROVAL LETTER

TO: Murchison, Charles

FROM: University of Alabama at Birmingham Institutional Review Board
Federalwide Assurance # FWA00005960
IORG Registration # IRB00000196 (IRB 01)
IORG Registration # IRB00000726 (IRB 02)

IORG Registration # IRB00012550 (IRB 03)

DATE: 05-Oct-2020

RE: IRB-300005964
Application of Discriminative and Generative Longitudinal Machine Learning Models in Multi-Study Alzheimer's Disease Datasets

---

The IRB reviewed and approved the Initial Application submitted on 31-Aug-2020 for the above referenced project. The review was conducted in accordance with UAB's Assurance of Compliance approved by the Department of Health and Human Services.

Type of Review:	Exempt
Exempt Categories:	4
Determination:	Exempt
Approval Date:	05-Oct-2020
Approval Period:	No Continuing Review

Please note:

- Please submit a Personnel Amendment to add the faculty advisor to your IRB Personnel eForm.

Documents Included in Review:

- IRB 300005964 Exemption Application
- IRB 300005964 Letter of Authorization
- IRB EPORTFOLIO
- IRB PERSONNEL EFORM