CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

MedBench: Benchmarking Machine Learning Methods for Medical Tabular Prediction

Master's thesis in Computer science and engineering

Alva Stöckel
Vidar Höök

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2026

Master's thesis 2026

MedBench: Benchmarking Machine Learning Methods for Medical Tabular Prediction

Alva Stöckel
Vidar Höök

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2026

Alva Stöckel
Vidar Höök

Supervisor: Fredrik Johansson, Computer Science and Engineering
Examiner: Gerardo Schneider, Computer Science and Engineering

Master's Thesis 2026
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2026

Alva Stöckel
Vidar Höök
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

Tabular data is central to medical machine learning, yet recent tabular foundation models have mostly been evaluated on general-purpose benchmarks. We evaluate whether such models transfer to medical tabular prediction by constructing a benchmark from the ADNI, MIMIC-IV, and NHANES datasets, covering prognostic and diagnostic tasks across binary classification, multiclass classification, and regression. TabPFN, TabICL, and MITRA are compared with classical baselines using modality-specific metrics, aggregate rank analyses, and bootstrap uncertainty estimates.

TabPFN and TabICL achieved the strongest overall performance on small- and medium-sized tasks, consistently ranking above classical baselines and tuned XGBoost. MITRA was competitive but less consistent. The advantage of TabPFN and TabICL was clearest on medium-sized tasks, while smaller tasks showed higher rank uncertainty due to limited sample sizes and fewer task instances. However, the largest tasks could not be evaluated with the full TabPFN and TabICL configurations under the available hardware constraints. On reduced-size variants of these tasks, TabPFN and TabICL remained competitive, but the results did not provide clear evidence that this advantage persisted as the number of samples increased.

The results indicate that tabular foundation models can transfer effectively to heterogeneous medical prediction tasks, especially in small- and medium-data regimes. However, their advantages are not uniform across dataset scales, and aggregate benchmark results can overstate practical usefulness if failed runs and subsampling are not reported. Future medical tabular benchmarks should therefore evaluate both predictive performance and computational feasibility.

This study used data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database.

Keywords: Computer, science, computer science, engineering, project, thesis.

Acknowledgements

We would like to thank our supervisor, Fredrik Johansson, for their guidance and support throughout this thesis. Their feedback and advice helped shape the direction of the project and improve the final work.

Data used in preparation of this thesis was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. As such, the investigators within ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this thesis. A complete listing of ADNI investigators can be found at `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`.

Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

The MIMIC-IV component of this study was adapted from Yet Another ICU Benchmark (YAIB). We thank the YAIB developers for making their benchmark framework and task definitions publicly available.

Data from MIMIC-IV was obtained through PhysioNet. We acknowledge the MIT Laboratory for Computational Physiology, Beth Israel Deaconess Medical Center, and the PhysioNet platform for making MIMIC-IV available for research.

Data from the National Health and Nutrition Examination Survey (NHANES) was obtained from the National Center for Health Statistics, Centers for Disease Control and Prevention.

Contents

List of Figures

List of Tables

1

Introduction

Over the past few years, pretrained foundation models, especially large language models (LLMs), have completely transformed machine learning, demonstrating great generalization capabilities through extensive pre-training on diverse datasets and enabling effective transfer to new tasks with minimal task-specific tuning. In contrast, progress in tabular models has been considerably slower, despite tabular data being one of the most prevalent format in many real-world domains [1]. Unlike images or text, tabular data lacks a natural spatial or sequential structure, and features vary widely in type, scale and meaning across datasets, making it difficult to design architectures with the strong inductive biases that have driven progress in other domains. This gap is particularly evident in healthcare, where most medically relevant information, including laboratory measurements, vital signs and medication, is stored as structured tabular data [2], and where the application of machine learning has long been limited by the need for task-specific feature engineering, retraining, tuning, and validation for each new clinical prediction problem. Medical prediction has for decades been a central aim in healthcare, as clinicians have sought to use patient data to anticipate disease onset, diagnose illness, guide treatment decisions, and allocate resources more effectively among others [3]. Yet most machine learning models in this domain remain task-specific, trained from scratch on individual datasets and unable to leverage knowledge from related clinical problems [4].

Recent advances in tabular machine learning, inspired by the success of foundation models in other domains, offer a potential path forward. A new generation of models aims to learn generalizable representations from tabular data, enabling transfer across tasks and datasets without extensive retraining [5]. This raises two central questions for medical machine learning:

1. Can tabular foundation models pretrained on synthetic data generalize to real-world medical datasets?

2. How does their performance compare to classical baselines across datasets of varying size and clinical context?

To answer these questions, benchmarks must reflect the breadth and complexity of clinically relevant tabular tasks. General-purpose tabular benchmarks are useful for measuring methodological progress, but they do not necessarily capture the data characteristics, outcome definitions, and evaluation priorities that arise in healthcare. Medical benchmarks therefore need standardized datasets, metrics, and evaluation

protocols that allow fair comparison across methods while also reflecting clinically or epidemiologically meaningful prediction settings.

Medical prediction can be categorized into several settings, including diagnostic tasks where the goal is to identify a condition already present, and prognostic tasks where the goal is to predict future events, disease progression, treatment response, or changes in clinical measurements. These tasks may involve binary, continuous, multiclass, or ordinal outcomes, and may be motivated by either clinical decision-making or epidemiological research. A representative benchmark must cover both, as model performance may differ substantially across these settings. General-purpose tabular benchmarks exist but are not designed with the specific challenges of healthcare in mind. Healthcare-specific benchmarks have been proposed but remain limited in scope. Existing works such as Harutyunyan et al. [6] and YAIB [7] are centered on ICU time-series settings and do not reflect the diversity of real-world tabular medical prediction tasks. Non-healthcare benchmarks are likewise insufficient, as medical datasets differ widely in variable sets, missingness patterns and patient populations, meaning that model performance on a single dataset provides limited insight into generalization across real-world settings.

Despite growing interest in tabular foundation models for healthcare, there is currently no standardized way to evaluate how these models perform across heterogeneous medical prediction tasks. Without a benchmark that incorporates this variability, it is not possible to assess whether the strong performance of tabular foundation models on general benchmarks extends to the heterogeneous conditions found in real-world medical prediction.

The goal for this project is therefore to create a benchmark suite of medical prediction tasks to enable the comparison of different tabular foundation models in the medical field. The benchmark spans three datasets covering complementary clinical settings: NHANES for population health, MIMIC-IV for acute ICU care, and ADNI for longitudinal disease progression, comprising 16 tasks spanning binary classification, multiclass classification, and regression across datasets ranging from a few hundred to several million observations. We evaluate TabPFN, TabICL, and MITRA against classical baselines, reporting model performance, aggregate rankings, and computational feasibility across tasks and datasets, with the aim of providing a reusable evaluation framework for future tabular foundation models in medical prediction.

2

Background

This chapter provides the background necessary to understand the benchmark design and experimental results. Tabular data refers to datasets organized as rows and columns, where each row represents an observation, such as a patient, and each column represents a feature or variable, such as age, blood pressure, or a laboratory measurement. The prediction task is to learn a mapping from these input features to a target variable, which may be a binary outcome, a class label, or a continuous value. A benchmark is a standardized collection of tasks, datasets, and evaluation protocols that enables reproducible comparison of methods under consistent conditions [8]. We review the challenges of learning from medical tabular data, existing benchmarks, and the two main classes of methods evaluated: classical machine learning models, which are trained from scratch on each task, and tabular foundation models, which leverage pre-training to generalize across tasks without task-specific retraining.

2.1 Challenges in Learning from Medical Tabular Data

Although tabular data is one of the most common data formats used in machine learning, it poses several challenges, including heterogeneous feature types, missing values, and limited cross-dataset generalization [9]. A key reason tabular data is challenging is that it lacks the regular spatial or sequential structure found in image and text data. In images, nearby pixels share information, and in text, words follow grammatical and semantic patterns. Without such structure, it is difficult to design architectures with strong inductive biases, and models must instead treat each feature independently [9].

Instead, tabular datasets often contain mixed-type columns, such as numerical, categorical, and datetime features. This heterogeneity makes it difficult to design models that can handle all variables consistently, especially when feature distributions and scales differ substantially across datasets [10]. The difficulty is further increased by the fact that the meaning of a feature is often tied to its contextual metadata, such as column names and dataset descriptions, rather than the observed values alone [10]. As a result, tabular learning often depends on dataset-specific preprocessing choices and does not transfer cleanly across domains.

A further challenge concerns the structural properties of tabular data itself. Unlike

in images or text, there is no spatial or sequential structure to guide models toward relevant information, meaning that irrelevant features must be identified without architectural support [11]. The relationship between features and the target is also often irregular and non-smooth, making it difficult for models that assume smooth or locally consistent patterns to perform well [11]. These properties are particularly problematic in medical settings, where datasets often vary across institutions, patient populations, and measurement practices, and where the set of relevant features may differ substantially across tasks.

Medical tabular data encompasses a broad range of structured data sources, including electronic health records and clinical registry variables from acute care settings, population health surveys capturing demographic and lifestyle factors, and longitudinal research cohorts tracking disease progression over time [12]. This data is used for tasks including diagnosis, prognosis, patient outcome prediction, and treatment support. In clinical prediction modeling, a fundamental distinction is made between diagnostic models, which estimate the probability that a disease or condition is already present at the time of prediction, and prognostic models, which estimate the risk of future clinical outcomes over a specified time horizon [13]. However, the practical value of these prediction models depends on the quality and structure of the underlying healthcare data, which are often heterogeneous and inconsistently recorded across institutions [14]. These challenges are compounded by several factors specific to medical settings, including domain shift, informative missingness, fairness concerns, and dataset size.

Domain shift occurs when shifts in patient demographics, disease prevalence, or measurement practices across institutions or time periods lead to substantially degraded model performance when deployed in a new setting [15]. This problem, known as dataset shift, takes many forms in healthcare. A particularly common variant is domain shift, where differences in data collection practices, patient populations, or coding conventions across institutions or time periods alter feature distributions and prevalence rates. Evaluating robustness to such shifts therefore remains relevant, as it indicates whether a model generalizes beyond the specific conditions under which the data was collected.

Informative missingness refers to cases where the absence of a measurement is itself predictive, because it carries information about a patient's condition or the clinical decision-making process. This arises because clinical decisions shape what gets recorded: for example, vital signs are measured more frequently for higher-risk patients, meaning that missing values carry information about patient status [16]. This is particularly relevant for machine learning because models that treat all missingness as noise therefore risk learning incorrect associations, which motivates explicit handling of informative missingness in benchmark design.

Dataset size is an ongoing issue in the medical field, as many clinically important events are rare and healthcare datasets are often highly imbalanced [17]. This makes it difficult for models to learn reliable decision boundaries, especially when the minority class contains the outcomes of greatest interest. In addition, standard classifiers often perform poorly under such skewed distributions because they are

optimized for overall accuracy rather than rare event detection [17].

Fairness has long been an issue in the medical field, where high quality data have historically been sourced from more affluent and homogeneous populations [18]. This bias propagates during both training and validation, leading to worse predictions for underrepresented groups.

2.2 Benchmarks for Tabular Learning

Benchmarks play an important role in machine learning because they provide a common basis for comparing methods under the same conditions. Without standardized datasets, metrics, and evaluation protocols, it becomes difficult to tell whether performance differences reflect real methodological progress or simply differences in preprocessing, tuning, or task setup. Benchmarks provide a shared framework for evaluating methods under consistent conditions, allowing fair and robust comparison to state-of-the-art approaches and helping researchers understand model strengths and weaknesses [8]. Benchmark repositories also improve reproducibility by providing standardized metadata, train-test splits, and shareable experimental results, making it easier to compare studies objectively [19]. Benchmarks also help reduce selection bias by replacing ad hoc dataset choices with a more carefully curated and representative set of tasks [20].

Several general-purpose tabular benchmarks have been proposed to evaluate model performance across diverse datasets. OpenML provides a curated platform with benchmark suites such as OpenML-CC18 [19], which standardize evaluation through machine-readable tasks, shared train-test splits, and rich dataset metadata to support reproducible comparisons. The AutoML Benchmark (AMLB) [20] extends this idea by systematically evaluating multiple AutoML frameworks across a large collection of public classification and regression tasks, while also reporting accuracy, runtime trade-offs, and framework failures.

In recent years, however, the tabular learning landscape has changed rapidly with the rise of tabular foundation models, which have shifted attention toward stronger priors, low-data performance, and broader evaluation criteria, including distributional behavior and robustness across datasets. This development has also motivated living benchmark systems such as TabArena [21], introduced in response to the limitations of static benchmarks that are not updated when new models, model versions, or methodological issues emerge. TabArena addresses this through a continuously maintained benchmark built from manually curated datasets and publicly available evaluation results, allowing the community to track model performance under standardized conditions over time.

Medical benchmarks for tabular data exist but are often limited to a single dataset or a single clinical setting. Harutyunyan et al. [6] introduced a standardized benchmark built on MIMIC-III with four prediction tasks, including in-hospital mortality, decompensation, length of stay, and phenotyping, which established an important foundation for reproducible evaluation in clinical machine learning. Yet the benchmark remains restricted to a single ICU dataset, limiting its ability to capture vari-

ation across institutions and data sources. Yet Another ICU Benchmark (YAIB) [7] extends this idea by harmonizing multiple ICU datasets into a single pipeline with shared tasks such as mortality, sepsis, acute kidney injury, kidney function, and length of stay, but it is still confined to ICU time-series data rather than broader static clinical tabular prediction.

No existing benchmark spans multiple heterogeneous medical datasets beyond the ICU setting or provides a systematic comparison of modern tabular foundation models across diverse medical prediction tasks. This leaves it unclear both how well these models generalize across clinical contexts and how they compare with established machine-learning approaches under a common evaluation protocol. This gap motivates the need for a benchmark that covers complementary medical settings, including both population health and longitudinal disease monitoring alongside acute care.

2.3 Tabular Machine Learning

Classical approaches to tabular prediction span linear models, tree-based ensembles, and neural networks, each with distinct inductive biases that interact differently with the structural properties of tabular data. Linear models such as logistic regression and ridge regression impose strong assumptions of linearity and feature independence, which limits their expressiveness on tasks with complex interactions but confers interpretability and robustness in low-data regimes.

Tree-based ensemble methods have long dominated predictive modelling on tabular data [11], [22]. Random forests combine many decision trees trained on bootstrap samples, reducing variance through averaging, while gradient boosting takes a sequential approach where each new tree corrects the errors of its predecessors. Modern implementations such as XGBoost [23] incorporate regularization and efficient histogram-based split finding. Several properties of tree-based models align naturally with tabular data: decision trees learn piecewise constant functions that capture irregular, non-smooth relationships without imposing smoothness constraints [11], and gradient boosted trees can internally handle missing values and strongly varying feature scales by learning appropriate split points [22].

Standard MLPs are flexible and can in principle approximate complex functions, but in practice they have not consistently matched tree-based methods on tabular data [22]. This has been attributed to the lack of inductive biases suited to tabular structure: unlike images or text, tabular features have no natural ordering or spatial proximity, making it difficult for neural architectures to exploit local structure [11].

A shared limitation of all classical approaches is that they must be trained from scratch on each new task, with no mechanism for leveraging knowledge from related datasets. This motivates the development of tabular foundation models.

2.3.1 Tabular Foundation Models

A recent line of work aims to build tabular foundation models, building on the idea of prior-data fitted networks (PFNs), where a transformer is meta-trained to approximate Bayesian posterior predictions by observing many small supervised tasks sampled from a prior over data-generating processes [24]. These foundation models perform supervised learning via in-context learning, a paradigm in which the model receives labeled training examples directly as input and produces predictions for new instances in a single forward pass, without any gradient updates or task-specific fine-tuning [24]. This is fundamentally different from classical machine learning, where a model is trained by iteratively updating its parameters. Instead, the model learns to predict by attending to the provided examples at inference time, effectively treating the training set as context [5], [24]. The most prominent example is TabPFN [5] which is trained on millions of small synthetic classification problems generated from Bayesian networks, learning to map a set of labeled training examples directly to class probabilities for new inputs in a single forward pass.

Subsequent variants such as Real-TabPFN [25], TabICL [26] and MITRA [27] extended this paradigm by incorporating real-world data in pre-training and relaxing the original constraints on dataset size, enabling application to larger and more heterogeneous tables. The original TabPFN was limited to datasets with at most 1,000 training samples and 100 features at inference time, a restriction that made it impractical for many real-world tasks.

While these models achieve strong performance on small benchmark datasets [21], inference remains computationally demanding at scale due to the quadratic complexity of self-attention over the training set. Furthermore, their ability to generalize across heterogeneous clinical datasets with differing feature sets and missingness patterns remains insufficiently understood, motivating the need for unified evaluation benchmarks.

3

Methods

We propose a benchmark for evaluating tabular machine learning models across diverse clinical prediction problems, comprising 16 tasks from three datasets. The datasets were selected to represent complementary clinical settings and vary in size, structure, and access requirements. NHANES is fully open and requires no data use agreement. MIMIC-IV and ADNI are open-credentialed, requiring registration and acceptance of a data use agreement but no study-specific data sharing agreement.

3.1 Task Selection

Tasks were selected to reflect the breadth of real-world medical tabular prediction across three complementary dimensions. First, tasks span both diagnostic settings, where the goal is to identify a condition already present, and prognostic settings, where the goal is to predict future clinical outcomes. Second, tasks vary in data structure: NHANES provides cross-sectional population data, with each individual contributing a single survey response, MIMIC-IV provides dense measurements from acute care stays, and ADNI provides longitudinal follow-up of the same individuals over years. Third, tasks differ substantially in cohort size, ranging from a few hundred participants in ADNI conversion to several million observations in MIMIC-IV, enabling evaluation across data regimes where different model families are expected to have different advantages. Finally, tasks cover binary classification, multiclass classification, and regression problems, reflecting the diversity of prediction targets encountered in clinical practice.

Other datasets were evaluated for inclusion but were excluded when they did not add a substantially distinct clinical setting, patient population, prediction task, or tabular feature space compared with the datasets already selected. The final selection prioritized datasets that together provide complementary clinical settings, cohort characteristics, task types, and dataset scales. MIMIC, NHANES, and ADNI were therefore chosen to balance breadth, relevance, accessibility, and reproducibility.

Tasks and outcome definitions are grounded in prior literature, replicating previous study setups as closely as possible. Where exact replication is not feasible, the closest available definitions are followed and deviations documented transparently, to maximize comparability, reproducibility, and interpretability. Detailed task specifications, including cohort definitions, feature lists, and summary statistics, are provided in Appendix A.

Table 3.1: Task overview with size and outcome statistics. For binary and multiclass tasks, outcome is reported as prevalence (Prev.) per class; for regression tasks, mean (μ) and standard deviation (σ) are reported.

Dataset	Task	Type	n (samples)	Outcome	Features
ADNI	AD conversion	BC	697	Prev. 36.7%	18 [28]
	Diagnosis (DX)	MC	5,609	24.3 / 43.9 / 31.8%	40 [29]
	ADAS-Cog 13	Reg	5,521	$\mu = 17.45, \sigma = 12.31$	40
	Ventricle vol.	Reg	3,529	$\mu = 43,426, \sigma = 23,407$	40
MIMIC-IV	Mortality	BC	49,526	Prev. 7.2%	249 [7]
	AKI	BC	2,706,946	Prev. 12.3%	249
	Sepsis	BC	2,433,728	Prev. 3.9%	249
	Kidney function	Reg	33,951	$\mu = 1.47, \sigma = 1.42$	249
	Length of stay	Reg	4,720,378	$\mu = 67.75, \sigma = 58.79$	249
NHANES	COPD	BC	39,742	Prev. 7.7%	41 [30]
	Depression	BC	26,392	Prev. 9.1%	16 [31]
	Hypertension	BC	51,929	Prev. 34.4%	7 [32]
	HbA1c	Reg	47,559	$\mu = 5.74, \sigma = 1.08$	21 [33]
	MASLD	BC	3,420	Prev. 30.1%	24 [34]
	NAFLD	BC	2,270	Prev. 9.4%	15 [35]
	Diabetes	MC	46,542	Prev. 14.5 / 34.3 / 51.2%	18 [36]

Abbreviations: BC = binary classification; MC = multiclass classification; Reg = regression; Prev. = prevalence

3.1.1 MIMIC

MIMIC-IV is a large, publicly available electronic health record dataset from Beth Israel Deaconess Medical Center and a prominent resource for clinical machine learning research [37], [38], [39]. It includes patient measurements, diagnoses, procedures, treatments, and clinical notes [37]. Its ICU module provides rich bedside data such as charted observations, intravenous infusions, and patient outputs, making it well suited for acute care prediction problems. MIMIC-IV was selected because it provides a large, well-established ICU-based setting with rich clinical data and public availability, supporting reproducible research [37]. In addition, the dataset has been the basis for extensive prior work, meaning that many clinically meaningful prediction tasks are already well established and can be adopted or adapted for benchmark construction [37], [38]. Access to MIMIC-IV requires completion of the PhysioNet credentialing process, acceptance of a data use agreement, and completion of the required CITI training.

Tasks were adopted from YAIB [7], which defines five prediction tasks developed in collaboration with clinicians: mortality, acute kidney injury, and sepsis as binary classification tasks, and kidney function and length of stay as regression tasks, providing coverage of both classification and regression settings within the ICU prediction domain.

ICU Mortality. ICU mortality is defined as death occurring the ICU stay, determined via the hospital expire flag in the MIMIC-IV admissions table, following YAIB [7]. ICU mortality is a prognostic binary task because it captures severe short-

term deterioration and is a standard endpoint in critical care. As MIMIC-IV does not distinguish between ICU and hospital mortality directly, ward transfer records were used to determine whether death occurred in the ICU.

Acute Kidney Injury. Acute kidney injury (AKI) is a prognostic binary task, defined as KDIGO stage ≥ 1, based on either an increase in serum creatinine or low urine output, following YAIB [7]. Baseline creatinine was defined as the lowest creatinine measurement over the preceding seven days.

Sepsis. Sepsis onset is a prognostic binary task defined using Sepsis-3 criteria, where sepsis is identified as organ dysfunction due to infection, following YAIB [7]. Organ dysfunction was defined as an increase in SOFA score 2 points compared to the lowest value over the preceding 24 hours. Suspicion of infection was defined as simultaneous use of antibiotics and culture of body fluids.

Kidney Function. Kidney function is a regression task, defined as the median serum creatinine level during the second day of ICU stay, following YAIB [7].

Remaining Length of Stay. Remaining length of stay is a regression task defined as the time in hours remaining until ICU discharge, calculated at each hour of the stay, following YAIB [7]. A maximum forecasting window of seven days was applied, as predictions beyond this interval were judged to be of limited clinical relevance.

3.1.2 NHANES

The National Health and Nutrition Examination Survey (NHANES) is a nationally representative U.S. survey that combines interviews, physical examinations, and laboratory measurements [40]. Since 1999, NHANES has operated as a continuous survey, with data released in biennial cycles [40] providing a rich and diverse set of variables suitable for constructing a wide range of chronic disease prediction tasks.

NHANES was selected for its breadth and open availability. As a nationally representative survey it also captures demographic diversity across the US population. Unlike many clinical datasets, NHANES requires no data use agreement, enabling full reproducibility. Tasks were selected based on reference studies in the clinical ML literature, with cycle inclusion varying per task depending on variable availability.

COPD. COPD is a diagnostic binary task defined from self-reported emphysema, chronic bronchitis, or COPD, following Wang et al. [41]. COPD is the fourth leading cause of death worldwide, responsible for approximately 5% of all global deaths, making early and accurate identification a clinical priority [42]. Features include demographic variables, smoking history, secondhand smoke exposure, comorbidities, dietary health status, and a range of laboratory biomarkers.

Depression. Depression is a diagnostic binary task, defined using the Patient Health Questionnaire (PHQ-9), with a score of ≥ 10 indicating moderate to severe

depression, following Vu et al. [31]. Depression affects an estimated 280 million people worldwide, corresponding to about 5% of adults, and is associated with substantial losses in productivity and quality of life [43]. The PHQ-9 is a widely used depression screening instrument with established validity for identifying major depression in population settings [31]. Features include demographic variables, poverty income ratio, BMI, blood pressure, eGFR, lifestyle factors, and cardiometabolic conditions.

Diabetes Staging. Diabetes staging is a multiclass classification task with three ordered classes: normoglycaemia, prediabetes, and diabetes, defined using American Diabetes Association (ADA) glycaemic thresholds [44]. Diabetes is defined as HbA1c $\geq$ 6.5% or fasting plasma glucose $\geq$ 126 mg/dL. Prediabetes is defined as HbA1c $\geq$ 5.7% or fasting plasma glucose $\geq$ 100 mg/dL. All remaining adults are classified as normoglycaemic [44]. Diabetes affects over 830 million adults worldwide and prevalence has nearly doubled since 1990, making early identification a global health priority [45]. The cohort is restricted to adults aged 18 and above with available HbA1c and fasting glucose measurements. Features include demographic variables, anthropometric measurements, blood pressure, lipid panel, smoking history, and socioeconomic indicators.

HbA1c. Glycated haemoglobin (HbA1c) is a regression task, as it is predicted as a continuous outcome following Ezz et al. [33]. HbA1c reflects average plasma glucose over the previous eight to twelve weeks and, unlike fasting glucose, requires no special patient preparation, making it a practical biomarker for both monitoring and diagnosis of diabetes [46]. Features include metabolic markers, anthropometric measurements, blood pressure, and laboratory values including serum glucose, lipids, and kidney function indicators.

Hypertension. Hypertension is a diagnostic binary task, defined as a measured systolic blood pressure of $\geq$ 130 mmHg, following López-Martínez et al. [32]. Hypertension affects an estimated 1.4 billion adults worldwide and is a major cause of premature death, yet nearly half of those affected are unaware of their condition [47]. Features include age, sex, race/ethnicity, BMI, smoking status, diabetes status, and chronic kidney disease (CKD). CKD is derived from self-reported kidney function for cycles prior to 2015–2016, and from estimated glomerular filtration rate (eGFR) and albumin-creatinine ratio for later cycles, following the approach used in López-Martínez et al. [32].

MASLD. MASLD was introduced in 2023 through a multisociety Delphi consensus process to replace the term NAFLD, defining hepatic steatosis in the presence of at least one cardiometabolic risk factor [48]. In this benchmark, MASLD is treated as a diagnostic binary task, defined as a controlled attenuation parameter (CAP) $\geq$ 302 dB/m measured by transient elastography, following Zhang et al. [34]. Individuals with hepatitis B or C infection or frequent alcohol consumption are excluded. Features include anthropometric measurements, blood pressure, lipid panel, liver function markers, and inflammatory indicators.

NAFLD. Non-alcoholic fatty liver disease (NAFLD) in adolescents is a diagnostic binary task, defined as alanine aminotransferase (ALT) $>$ 26 IU/L in males and $>$ 22 IU/L in females, following Zhang et al. [35]. NAFLD is a leading cause of liver disease worldwide, with a global adult prevalence of 32% and a steadily increasing burden over time [49]. Features include anthropometric measures, lipid panel, blood count, and metabolic markers. The cohort is restricted to adolescents aged 12-19 years, consistent with the reference study's focus on youth populations.

3.1.3 ADNI

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal, multicenter observational study designed to validate biomarkers for Alzheimer's disease clinical trials [50]. ADNI data include clinical, imaging, genetic, and other biomarker measurements collected across repeated visits [50], making it well suited for modeling disease progression. Access to ADNI requires a formal application and acceptance of a data use agreement. Data was obtained through the Image and Data Archive (IDA) using the archived `ADNIMERGE` CSV file, which corresponds to the data source used by the reference studies this benchmark aims to reproduce. ADNI was selected to add a longitudinal, disease specific perspective to the benchmark, complementing the population-level setting of NHANES and the acute care setting of MIMIC-IV.

TADPOLE. Three tasks were motivated by the setup of the TADPOLE Challenge [29], which defined standardized prediction tasks for Alzheimer's disease progression using ADNI data. Each task predicts an outcome 12 months ahead from a given visit, deviating from the original challenge which did not specify a single fixed prediction horizon. The three outcomes are diagnosis (DX, multiclass), ADAS-Cog 13 score (regression), and ventricular volume (regression).

- **Diagnosis (DX):** Multiclass classification into cognitively normal, MCI, or dementia.
- **ADAS-Cog 13:** Regression of the 13-item Alzheimer's Disease Assessment Scale score, a continuous measure of cognitive impairment severity.
- **Ventricular volume:** Regression of ventricular volume, a structural biomarker of brain atrophy associated with Alzheimer's disease progression.

AD Conversion. AD conversion is a binary classification task, defined as conversion from mild cognitive impairment (MCI) to Alzheimer's disease dementia within the follow-up period, following Grassi et al. [28]. A subject is classified as a converter if a dementia diagnosis is recorded and MMSE score achieved at any follow-up visit after baseline. The cohort is restricted to MCI subjects at baseline with at least 36 months of follow-up. Features include sociodemographic characteristics, clinical scale scores, and neuropsychological test scores.

4

Experimental Setup

This chapter describes the experimental setup used to evaluate the benchmark. The benchmark aims to assess whether tabular foundation models, which leverage pretraining to generalize across tasks, can match or outperform classical machine learning models across heterogeneous medical prediction settings. To this end, we evaluate a set of classical models and tabular foundation models across the 16 tasks defined in Chapter 3, spanning three datasets, three clinical settings, and three prediction modalities. A shared preprocessing scheme, consistent evaluation metrics, and dataset-specific train-test split strategies are applied throughout to ensure comparability across tasks and models.

4.1 Preprocessing

To make the benchmark compatible with a broad range of tabular prediction models, we applied a common preprocessing scheme to the generated cohort tables. Continuous variables were imputed using the training-set mean, while categorical variables were imputed using the most frequent category in the training set. In addition, missingness-indicator features were added to preserve information about whether a value was originally missing.

This preprocessing strategy was chosen to provide a simple and consistent baseline across tasks while avoiding the additional computational cost of evaluating multiple imputation strategies. It also makes the resulting data representation compatible with models that do not natively handle missing values. MITRA was treated as an exception to this preprocessing scheme. For MITRA, the raw cohort data frames were provided directly, since this is the input format expected by the model. Dataset-specific preprocessing and splitting procedures are described below.

ADNI tasks are longitudinal, with each sample defined by a participant at an index visit and a prediction target derived from a subsequent follow-up visit. Splitting was performed at the participant level to prevent data leakage across splits, with stratification applied to preserve outcome distributions.

MIMIC-IV time-series observations were converted into fixed-length feature vectors following the tabularization approach used in YAIB. For each continuous variable, five derived features were generated: the last observed value, a missingness indicator, and the maximum, minimum, and mean over the stay. One deviation from YAIB

was the omission of the per-feature time vector, which contained repeated rather than feature-specific time information; this was replaced with a single time feature. For kidney function and mortality tasks, only the final sample per stay was used. Splitting was stratified at the stay level.

Unlike ADNI and MIMIC-IV, the NHANES tasks were evaluated using a cycle-based temporal holdout split rather than a random train-test split. Earlier survey cycles were used for training, and later cycles were used for testing. This split was chosen to evaluate model performance under temporal distribution shift, since NHANES cohorts may differ across survey years because of changes in population characteristics, measurement procedures, and survey design.

4.2 Models

Models were selected to represent a range of approaches to tabular prediction, and were evaluated under a primarily out-of-the-box setting to reflect the performance a user could reasonably obtain without extensive task-specific tuning. A majority classifier dummy baseline is included to provide a lower bound on performance. Logistic regression, MLP, and random forest represent classical machine learning approaches, while XGBoost[23] represents modern gradient boosted trees. TabPFN [5] and TabICL [26] represent the tabular foundation model paradigm, enabling evaluation of in-context learning approaches against classical baselines. MITRA[27] is also included as an additional tabular foundation model. Unlike the other models, MITRA receives raw cohort data directly rather than preprocessed feature vectors, as this is the input format expected by the model.

All models except XGBoost were evaluated using default hyperparameters. For the tabular foundation models, configuration changes were limited to settings required for successful execution under the available hardware constraints, and were treated as implementation constraints rather than task-specific optimization. XGBoost was evaluated in two configurations: an untuned variant using default hyperparameters, and a tuned variant where hyperparameters were optimized using Optuna [51], providing a stronger classical baseline for comparison with the foundation-model-style methods. Full hyperparameter details and model versions are provided in Appendix.

4.3 Evaluation Metrics

The primary evaluation metric for binary classification tasks is the area under the receiver operating characteristic curve (AUROC), which measures discriminative ability independently of classification threshold, consistent with established practice in clinical ML benchmarking [6], [7]. For multiclass tasks, macro-averaged AUROC is used. For regression tasks, root mean squared error (RMSE) (FIXME: use nrmse in plots) is used as the primary metric, following the convention adopted in tabular benchmarking [21]. An additional column was later added to the heatmaps in Figures 5.1 and 5.2

5

Results

5.1 Main Performance Comparison

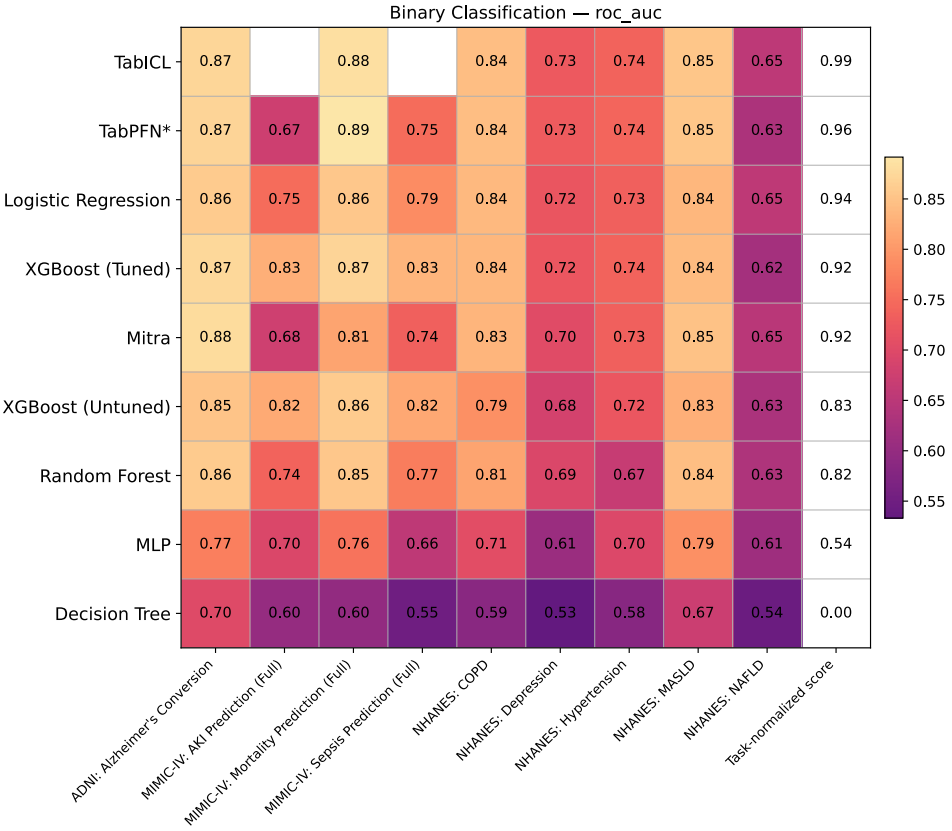
Across the benchmark, as seen in figures 5.1, 5.2, the strongest overall performance was obtained by the tabular foundation models. TabPFN and TabICL were consistently among the highest-ranked models across task modalities, and both generally outperformed the classical machine-learning baselines, including the Optuna-tuned XGBoost variant. MITRA also performed competitively and was often among the better-performing models, although it did not match the consistency of TabPFN and TabICL. This result suggests that recent tabular foundation models can transfer to medical prediction tasks, despite being evaluated on heterogeneous datasets with different cohort definitions, target types, and missingness patterns. The performance advantage was not limited to a single modality; the same general pattern was observed across the main task groups, indicating that the result is not explained only by one subset of the benchmark.

However, the strong performance of tabular foundation models came with important computational limitations. The largest benchmark datasets, corresponding to the three largest MIMIC-IV tasks, could not be evaluated with the full foundation-model configurations under the available hardware constraints. TabICL did not finish on these tasks, while TabPFN required subsampling for large-scale tasks such as sepsis, acute kidney injury, and length of stay.

Overall, the main performance comparison shows that TabPFN and TabICL achieved the strongest results on the tasks where the full models could be evaluated, but their advantage should be interpreted together with their computational requirements and their behavior on large-scale tasks. The results therefore support the predictive usefulness of tabular foundation models for many medical benchmark tasks, while also showing that their performance advantage is not uniform across dataset scales.

Figure 5.3 reports aggregate bootstrap ranks for the completed full-model evaluations. TabPFN and TabICL have the lowest mean aggregate ranks across the evaluated modalities, and their bootstrap intervals remain among the lowest-ranked models. This indicates that their relative ranking is stable under resampling for the set of tasks included in the bootstrap analysis.

The interpretation of this figure is limited by the task subset included. The largest MIMIC-IV tasks, as well as the reduced-size large-task variants, are not included in



*TabPFN was evaluated using the subsampled setting for Sepsis and AKI.

Figure 5.1: Benchmark performance for binary classification tasks. Task-normalized score excludes all Sepsis and AKI tasks. Sorted by task-normalized score (AUROC)

this aggregate bootstrap ranking. Consequently, the figure summarizes uncertainty for the small- and medium-sized benchmark tasks, but does not address model behavior on the largest datasets.

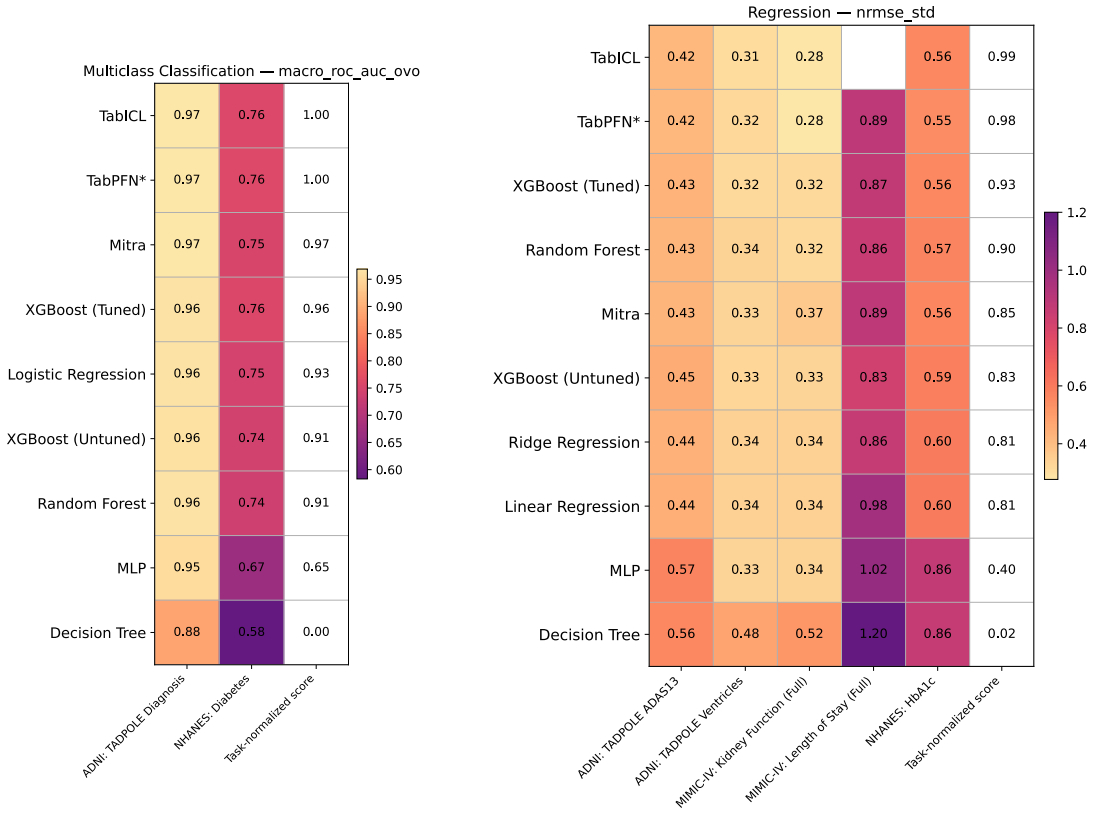
5.2 Robustness and Uncertainty

5.3 Dataset-Specific Results

5.3.1 ADNI

Figure 5.4 shows wider rank intervals than the full aggregate analysis, consistent with the smaller number of ADNI tasks and their limited sample sizes. The overall pattern remains similar: foundation-model-style methods are among the best-ranked models, but the separation between TabPFN, TabICL, and MITRA is small. In this subset, MITRA ranks approximately on par with the other foundation-model-style methods.

Since the ADNI tasks are all small prognostic longitudinal prediction tasks, with sample sizes from a few hundred to a few thousand observations, small rank differ-



Multiclass classification Regression

*TabPFN was evaluated using the subsampled setting for LOS.

Figure 5.2: Benchmark performance across Multiclass and Regression. For nrmse, lower is better. Task-normalized score excludes Length of Stay tasks.

ences should be interpreted cautiously.

5.3.2 MIMIC-IV

Figure 5.5 reports aggregate bootstrap ranks for the two smaller MIMIC-IV tasks completed by all models. Most models have overlapping rank intervals, indicating substantial uncertainty in their relative ordering. TabPFN and TabICL are the exception, remaining consistently in the two best aggregate ranks across bootstrap replicates.

Because this subset contains only two tasks, the result should not be interpreted as a general trend for all MIMIC-IV tasks. It indicates that, on the smaller MIMIC-IV tasks included in this analysis, TabPFN and TabICL were consistently ranked above the other methods. MITRA ranks near the lower end of the model set in this subset, above only the decision tree baseline, but this should be interpreted cautiously given the small number of tasks.

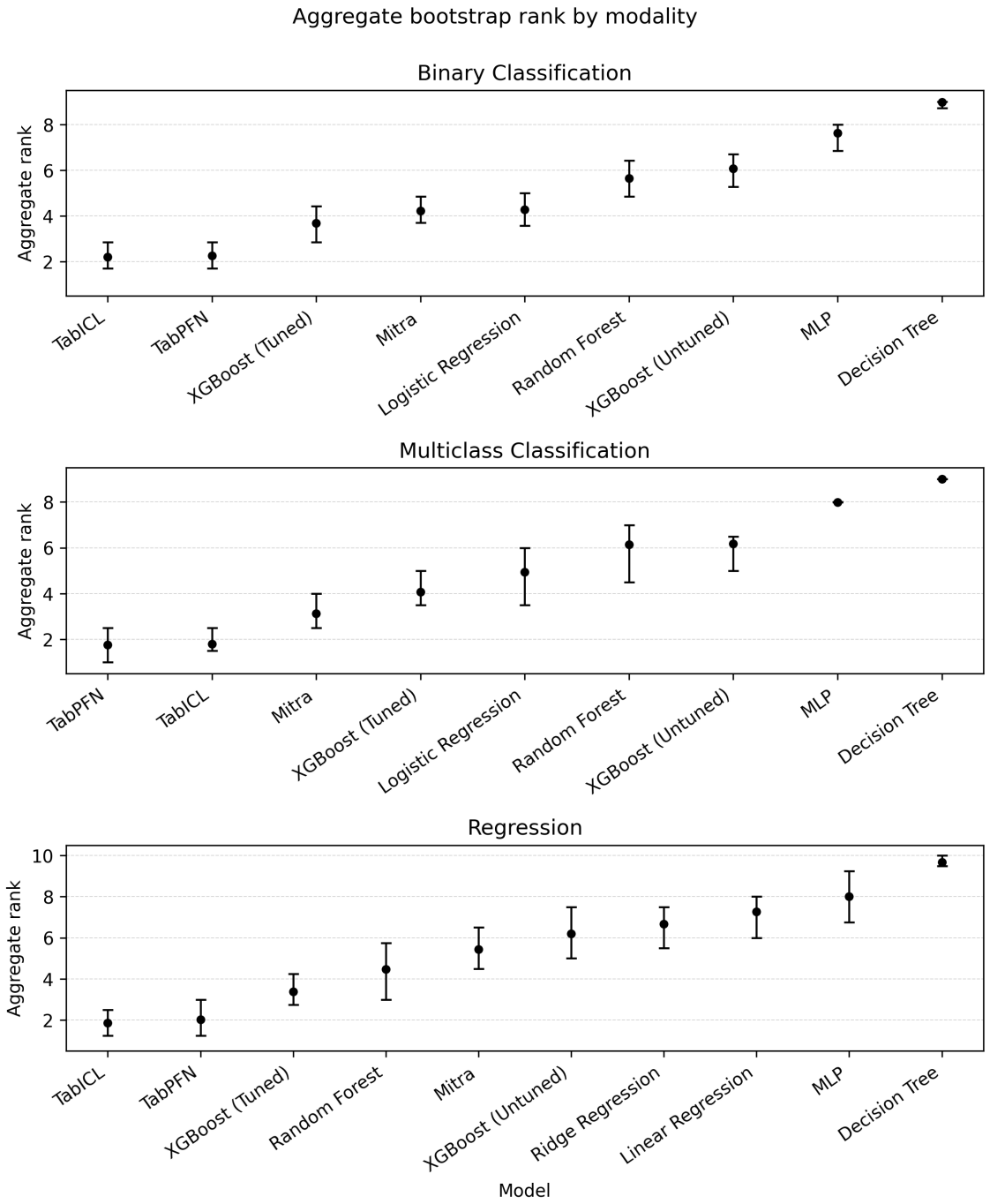


Figure 5.3: Aggregate bootstrap ranks by modality. Models were ranked independently within each task using the modality-specific primary metric: ROC-AUC for binary classification, macro-auroc-ovo for multiclass classification, and normalized root-mean-squared error for regression. For each bootstrap replicate, models were ranked using replicate-level metric values, and task-level ranks were averaged across tasks with equal task weighting. Points show mean aggregate rank; intervals show the 2.5–97.5 percentile bootstrap rank interval. Lower values indicate better rank. Does not include the Sepsis, AKI, or Length of Stay tasks.

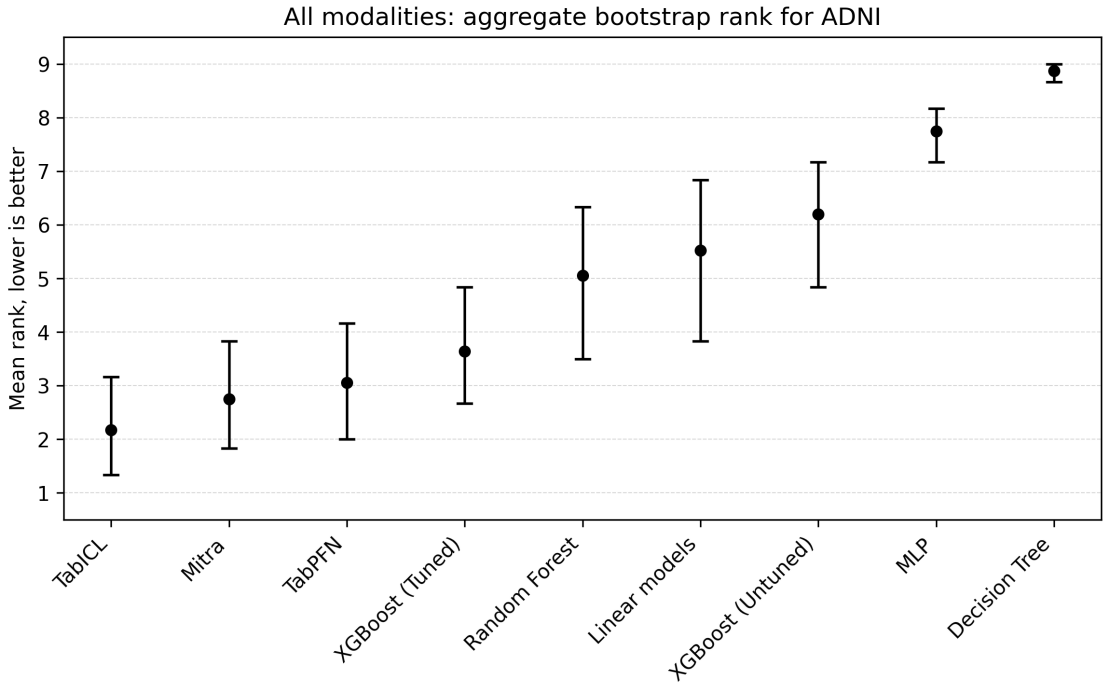


Figure 5.4: Aggregate bootstrap ranks for the ADNI tasks. Models were ranked within each task using the modality-specific primary metric, and ranks were averaged across tasks with equal task weighting. Points show mean rank and intervals show the 2.5–97.5 percentile bootstrap interval. Lower rank is better. Ridge regression and Logistic regression were grouped, and linear regression discarded.

To further investigate the behavior of the largest MIMIC-IV tasks, we evaluated reduced-size variants of these tasks at three dataset scales: tiny, small, and medium (see: Figure 5.6). For each variant, models were ranked using the same task-level ranking procedure as in the main analysis, and the resulting ranks were plotted as a function of dataset size (see: Figure 5.7).

The expected pattern was that the relative rank of TabPFN and TabICL would worsen as dataset size increased, indicating that their advantage is strongest in smaller-data regimes. The observed results provide only limited support for this pattern. The foundation-model-style methods do not show a consistent monotonic degradation across all variants. However, on the largest reduced-size setting, the default variant of XGBoost begins to rank above TabPFN and TabICL, while the tuned XGBoost model remains the best-ranked method across all evaluated sizes.

Overall, the size-scaling analysis suggests a possible weakening of the foundation-model advantage as the large MIMIC-IV tasks increase in size, but the evidence is not conclusive. The results are better interpreted as an indication that the large MIMIC-IV tasks differ from the smaller benchmark tasks, rather than as proof of a general monotonic relationship between dataset size and foundation-model rank.

Figure 5.8 shows relatively narrow bootstrap rank intervals compared with the ADNI and MIMIC-IV subset analyses. This likely reflects the larger number of NHANES

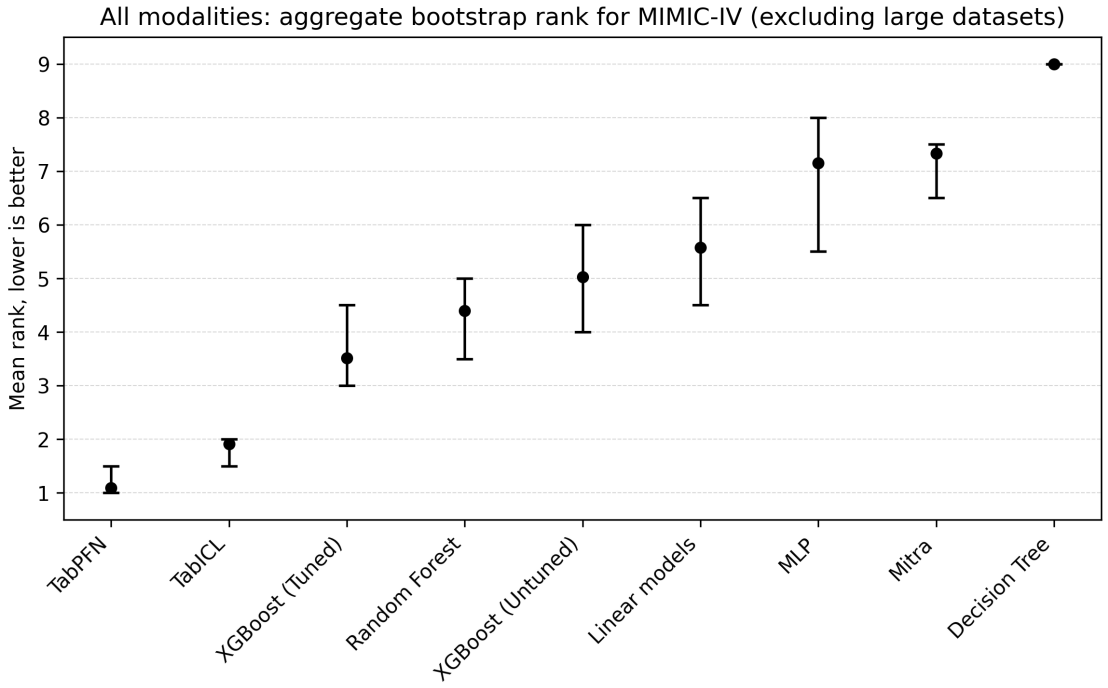
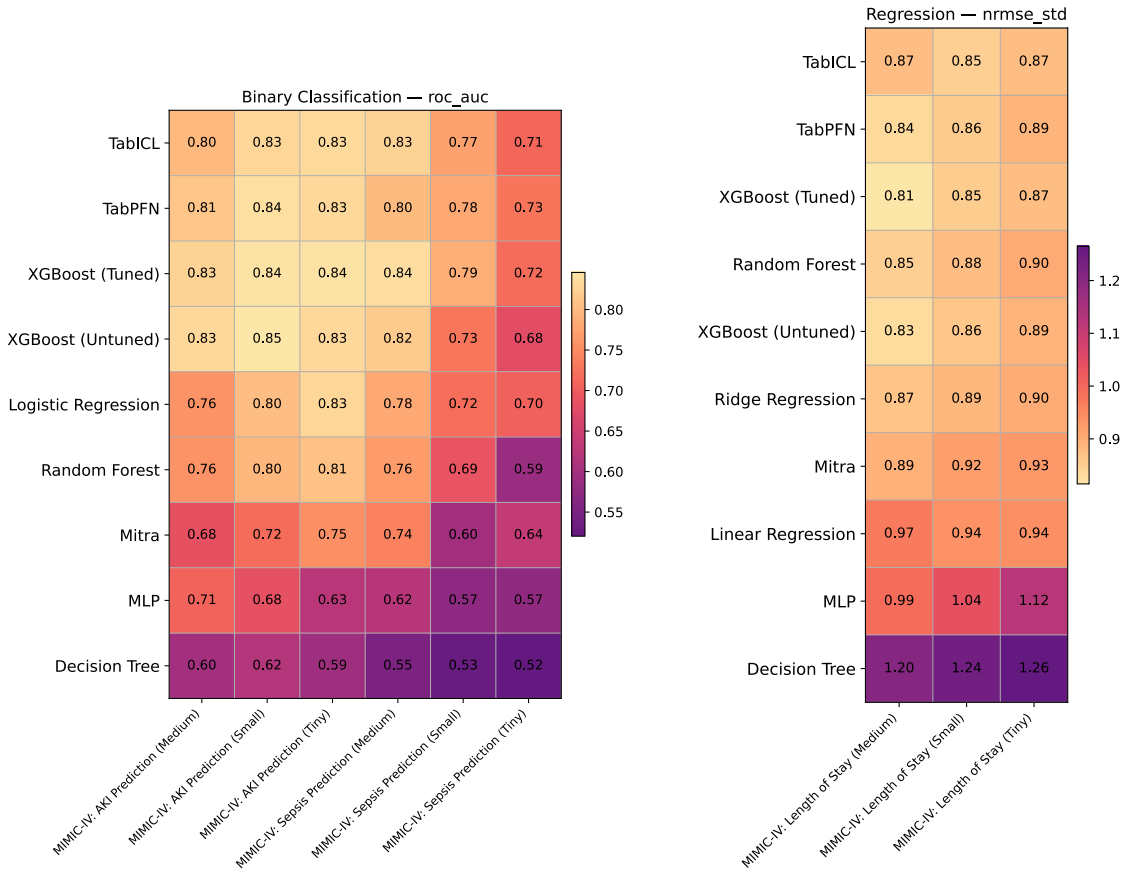


Figure 5.5: Aggregate bootstrap ranks for the two smaller MIMIC-IV tasks completed by all models. Models were ranked within each task using the modality-specific primary metric, and ranks were averaged across tasks with equal task weighting. Points show mean rank and intervals show the 2.5–97.5 percentile bootstrap interval. Lower rank is better. Ridge regression and Logistic regression were grouped, and linear regression discarded.

tasks and their larger sample sizes, ranging from the low thousands to the mid tens of thousands of observations.

The model ordering is stable across bootstrap replicates. TabPFN and TabICL have the two lowest aggregate ranks, with TabPFN ranking ahead of TabICL in this subset. Most intervals show limited overlap, suggesting that the relative ordering of models is more clearly separated for NHANES than for the smaller dataset-specific analyses.



Multiclass classification Regression

*TabPFN was evaluated using the subsampled setting for LOS.

Figure 5.6: Benchmark performance across Multiclass and Regression. For nrmse, lower is better. Task-normalized score excludes Length of Stay tasks. Sorted by task-normalized score (macro-auroc one-vs-one for multiclass, normalized (standard deviation) rmse for regression)

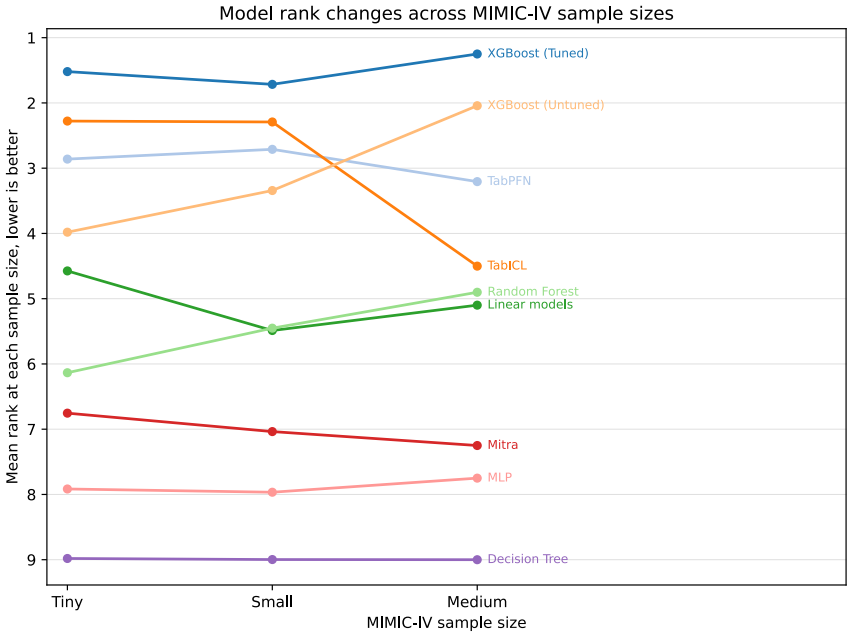


Figure 5.7: Aggregate bootstrap ranks for the two smaller MIMIC-IV tasks completed by all models. Models were ranked within each task using the modality-specific primary metric, and ranks were averaged across tasks with equal task weighting. Points show mean rank and intervals show the 2.5–97.5 percentile bootstrap interval. Lower rank is better. Ridge regression and Logistic regression were grouped, and linear regression discarded. Here, Tiny corresponds to around 50K samples, small around 150K samples, and medium around 500K samples.

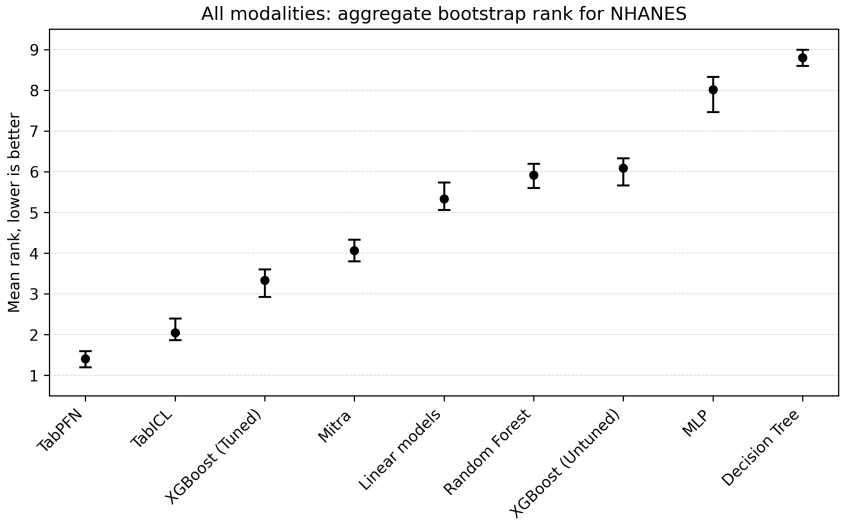


Figure 5.8: Aggregate bootstrap ranks for the NHANES tasks. Models were ranked within each task using the modality-specific primary metric, and ranks were averaged across tasks with equal task weighting. Points show mean rank and intervals show the 2.5–97.5 percentile bootstrap interval. Lower rank is better. Ridge regression and Logistic regression were grouped, and linear regression discarded.

6

Discussion

6.1 Generalization and practical trade-offs

With respect to generalization to real-world medical tabular data, the results indicate that tabular foundation-model-style methods can provide strong predictive performance across heterogeneous medical benchmark tasks. TabPFN and TabICL were consistently among the highest-ranked models across the evaluated task modalities, and their relative ranking was stable in the aggregate bootstrap analysis for the subset of tasks where full-model evaluation was completed. This suggests that prior-data-fitted and in-context tabular models are not limited to general-purpose tabular benchmarks, but may also be useful as strong general-purpose predictors in medical settings, particularly when task-specific model engineering is limited.

However, the predictive advantage of these models comes with a less favorable interpretability profile. In clinical prediction settings, model performance is rarely the only relevant criterion. A model may achieve higher AUROC or lower error while still being difficult to inspect, validate mechanistically, or explain to clinicians and other stakeholders. This is especially relevant for tabular foundation models, where the prediction is produced by a learned inference procedure rather than by a transparent set of fitted coefficients or an easily summarized tree ensemble. As a result, the practical value of the observed performance gains depends on whether the improvement is large enough to justify the additional opacity.

This trade-off is important because several of the observed differences are expressed through ranks rather than large, directly interpretable absolute performance gaps. Aggregate ranks are useful for comparing models across heterogeneous tasks and metrics, but they do not by themselves show whether a difference is clinically meaningful. A small performance improvement may be sufficient in some high-impact applications, but in other settings it may not compensate for reduced transparency, increased validation burden, or more difficult model governance. Therefore, the results should not be interpreted as implying that tabular foundation models are automatically preferable for medical deployment. Rather, they suggest that these models are promising candidates when predictive performance is the primary objective, while further analysis would be required to determine whether their gains remain worthwhile under clinical requirements such as calibration, subgroup reliability, auditability, and interpretability.

A practical interpretation is that tabular foundation models may be most useful as high-performing benchmark models or as components in early-stage model development. They can establish a strong performance reference point and may reduce the need for extensive task-specific tuning. For clinical use, however, their adoption would likely require additional explanation methods, calibration analysis, and external validation. In this sense, the present results support their predictive potential, but they do not remove the need to evaluate whether the resulting models can be trusted, understood, and maintained in a medical decision-making context.

6.2 Comparison with classical baselines across dataset scales

The comparison with classical baselines shows that the advantage of tabular foundation-model-style methods is conditional on dataset scale. Although TabPFN and TabICL achieved the strongest overall results across the completed benchmark tasks, this pattern was clearest in the small- and medium-sized tasks and became less consistent in larger sample regimes. In the reduced-size MIMIC-IV scaling analysis, the evidence did not support a simple monotonic degradation of TabPFN or TabICL as dataset size increased. However, the largest reduced-size setting showed stronger performance from XGBoost, with the default XGBoost variant ranking above TabPFN and TabICL, while the tuned XGBoost model remained the best-ranked method across the evaluated sizes. This suggests that the relative advantage of tabular foundation models depends strongly on dataset regime rather than being uniform across medical prediction tasks.

This finding is consistent with the practical role of gradient-boosted tree models in tabular learning. Methods such as XGBoost are mature, scalable, and well-suited to large structured datasets. They can benefit directly from increasing sample size, and their training and inference procedures are well understood. In contrast, tabular foundation-model-style methods are designed to transfer across tasks and may be especially attractive when the available task-specific data is limited, but this does not necessarily imply superiority when sufficient training data is available for a strong supervised baseline. In larger datasets, the benefit of learned prior information may become less important relative to the ability of classical supervised models to fit the task directly.

The large-scale setting also raises practical deployment considerations. While the present results do not provide conclusive evidence about performance at the full scale of the largest MIMIC-IV tasks, they do show that evaluation and deployment feasibility cannot be separated from model comparison. TabPFN required subsampling for large-scale tasks such as sepsis, acute kidney injury, and length of stay, while TabICL did not finish on the largest tasks under the available hardware constraints. In a large-scale clinical setting, running these models without subsampling could require very large amounts of accelerator or system memory, potentially reaching hundreds of gigabytes or more depending on the configuration and dataset size. Even if such resources are available, repeated evaluation, validation, and monitoring

could become prohibitively expensive.

For this reason, the comparison between foundation-model-style methods and classical baselines should be interpreted in terms of both predictive performance and operational feasibility. Tabular foundation models appear highly competitive on many small- and medium-sized benchmark tasks, but tuned classical models remain difficult to dismiss when larger datasets are available. In practice, XGBoost and related methods may offer a more favorable balance between performance, scalability, interpretability tooling, and deployment cost in large-sample regimes. The results therefore suggest a complementary rather than replacement-based view: tabular foundation models are strong additions to the medical tabular modelling toolkit, but classical models remain important baselines and may remain preferable in large-scale applications.

6.3 Limitations

The benchmark includes a limited number of datasets. The analysis was restricted to ADNI, MIMIC-IV, and NHANES for reasons of time, accessibility, and reproducibility. These datasets cover different medical settings and task types, but they do not represent the full range of available clinical tabular prediction problems. The conclusions should therefore be interpreted as evidence from a selected set of medical benchmarks rather than as a comprehensive evaluation of tabular foundation models across medicine.

The number and diversity of tasks is also limited. This is especially true for multiclass classification and, to a lesser extent, regression, where the benchmark contains substantially fewer tasks than for binary classification. This imbalance is not unique to this work. TabArena, a general-purpose benchmark for tabular machine learning, contains eight multiclass tasks out of a total of 51 [21], reflecting the relative scarcity of well-established multiclass prediction problems in the broader tabular learning literature. In the medical domain, this imbalance is further amplified, as many clinically relevant prediction problems are naturally binary or continuous, with comparatively few widely adopted multiclass prediction tasks. Consequently, the limited number of multiclass tasks restricts the strength of conclusions that can be drawn for this modality.

A further limitation concerns the extent to which the reproduced tasks correspond to meaningful clinical decision problems. The tasks are representative of common medical machine-learning benchmarks, but this does not necessarily mean that they address clinically relevant questions. Many published prediction tasks are designed primarily around convenient labels and standard performance metrics rather than around clearly defined clinical decisions. As a result, some tasks may predict outcomes or measurements that are not themselves useful decision targets, or that could be obtained more directly through existing clinical procedures. This limits the practical interpretation of high predictive performance. Strong benchmark results do not necessarily imply clinical usefulness unless the target, timing, input variables, and intended decision context are clinically justified.

Relatedly, the evaluation follows the metrics used in the reference studies, with AUROC serving as the main metric for binary classification. AUROC is useful for threshold-independent discrimination, but it is insufficient as a measure of clinical utility. It does not assess calibration, decision thresholds, net benefit, treatment consequences, or whether the model improves decision-making compared with existing clinical practice. This is particularly important in medical prediction, where the cost of false positives and false negatives may be highly asymmetric and context-dependent. Therefore, the present results should be understood as a comparison of predictive discrimination and error metrics, not as a demonstration of deployable clinical utility.

The benchmark also does not evaluate survival analysis or time-series data directly. Although some source datasets, especially electronic health record (EHR) datasets, contain longitudinal and time-dependent information, this data was represented as static tabular prediction tasks. This design makes the benchmark compatible with the evaluated tabular models, but it removes parts of the temporal structure that may be clinically important. Many medical problems are naturally time-to-event, recurrent-event, or sequential prediction problems, and EHR data often contains irregular measurements, changing treatments, repeated admissions, and evolving patient states. Methods designed for static tabular inputs may not capture these aspects adequately.

Hardware constraints also affected the evaluation. The largest MIMIC-IV tasks could not be evaluated with all foundation-model-style methods under their full configurations: TabPFN required subsampling, and TabICL did not complete on the largest tasks. This limits conclusions about their behavior at full clinical dataset scale. In deployment or repeated-validation settings, their memory and compute requirements may become a substantial barrier, particularly if full-scale evaluation requires hundreds of gigabytes or more of accelerator or system memory.

Finally, the study is also limited by the geographic and healthcare-system context of the selected datasets. The datasets used here are American, and they reflect patient populations, measurement practices, coding systems, access patterns, and healthcare structures that may not transfer directly to other countries or healthcare systems. As a result, performance on these datasets should not be assumed to generalize to other populations without external validation.

6.4 Future Work

6.4.1 Broader dataset and clinical-domain coverage

Future work should expand the benchmark to include a larger and more diverse set of medical datasets. The present study was limited to three data sources, which provided variation in cohort type and task setting but cannot represent the full range of medical tabular prediction problems. Additional datasets should include more disease-specific cohorts, primary-care data, specialist registries, imaging-derived tabular features, laboratory-focused datasets, and multi-institutional clinical datasets.

This would make it possible to evaluate whether the observed performance patterns hold across different diseases, measurement practices, patient populations, and clinical workflows.

6.4.2 External validation across populations and healthcare systems

Future work should include external validation cohorts. This is particularly important because medical prediction models are often sensitive to changes in population demographics, clinical practice, measurement routines, coding standards, and healthcare-system structure. A model that performs well on one dataset may not remain reliable when applied to another hospital, country, time period, or patient population.

External validation should therefore be performed across independent sites and, where possible, across healthcare systems outside the United States. This would help determine whether the relative performance of tabular foundation models transfers beyond the datasets used for model selection and benchmarking. Such validation would also allow future studies to evaluate temporal and geographic distribution shift, which is central to assessing whether benchmark performance is likely to translate into real-world clinical robustness.

6.4.3 More outcome types

Future work should broaden the set of evaluated outcome types. Although the benchmark included both classification and regression tasks, binary classification was the most extensively represented setting, and pure multiclass classification was represented by only a small number of tasks.

Future benchmarks do not need an equal or proportional distribution of outcome types, but should include enough tasks within each modality to support meaningful modality-specific evaluation. This would make it easier to assess whether model performance is stable across different prediction structures, including continuous outcomes, severity grades, functional scores, risk stages, and treatment response categories.

6.4.4 Survival analysis and longitudinal clinical data

Future work should evaluate survival analysis and longitudinal modelling directly. This includes censored outcomes, recurrent events, and longitudinal EHR trajectories, either by adapting tabular foundation models to these settings or by comparing them with methods specifically designed for survival analysis and time-series prediction. Such evaluations would help determine whether the observed performance patterns extend beyond static tabular endpoints.

6.4.5 Full-scale evaluation and compute constraints

Future work should perform full-scale evaluations under realistic compute constraints. In the present study, some large tasks could not be evaluated with all full model configurations. Future studies should investigate how tabular foundation models behave when applied to full clinical datasets, while also reporting memory usage, runtime, inference cost, and practical deployment requirements.

Bibliography

[1] G. Zabërgja, A. Kadra, C. M. M. Frey, and J. Grabocka, *Tabular data: Is deep learning all you need?* 2025. arXiv: 2402.03970 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2402.03970.

[2] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, pp. 1347–1358, Apr. 2019. DOI: 10.1056/NEJMra1814259.

[3] Z. Obermeyer and E. Emanuel, "Predicting the future big data, machine learning, and clinical medicine," *The New England journal of medicine*, vol. 375, pp. 1216–1219, Sep. 2016. DOI: 10.1056/NEJMp1606181.

[4] T. K. Tran, M. C. Tran, A. Joseph, P. A. Phan, V. Grau, and A. D. Farmery, "A systematic review of machine learning models for management, prediction and classification of ards," *Respiratory Research*, vol. 25, no. 1, p. 232, 2024, ISSN: 1465-993X. DOI: 10.1186/s12931-024-02834-x. [Online]. Available: https://doi.org/10.1186/s12931-024-02834-x.

[5] N. Hollmann et al., "Accurate predictions on small data with a tabular foundation model," *Nature*, vol. 637, pp. 319–326, Jan. 2025. DOI: 10.1038/s41586-024-08328-6.

[6] H. Harutyunyan, H. Khachatrian, D. Kale, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific Data*, vol. 6, Jun. 2019. DOI: 10.1038/s41597-019-0103-9.

[7] R. van de Water, H. Schmidt, P. Elbers, P. Thoral, B. Arnrich, and P. Rockenschaub, "Yet another icu benchmark: A flexible multi-center framework for clinical ml," 2024. arXiv: 2306.05109 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2306.05109.

[8] P. Orzechowski and J. H. Moore, "Generative and reproducible benchmarks for comprehensive evaluation of machine learning classifiers," *Science Advances*, vol. 8, no. eabl4747, 2022. DOI: 10.1126/sciadv.abl4747.

[9] W. Ren, T. Zhao, Y. Huang, and V. Honavar, *Deep learning within tabular data: Foundations, challenges, advances and future directions*, 2025. arXiv: 2501.03540 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2501.03540.

[10] B. van Breugel and M. van der Schaar, *Why tabular foundation models should be a research priority*, 2024. arXiv: 2405.01147 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2405.01147.

[11] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?," 2022. arXiv: 2207.08815 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2207.08815.

[12] Z. Wang, C. Gao, C. Xiao, and J. Sun, *Meditab: Scaling medical tabular data predictors via data consolidation, enrichment, and refinement*, 2024. arXiv: 2305.12081 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2305.12081.

[13] M. van Smeden, J. B. Reitsma, R. D. Riley, G. S. Collins, and K. G. M. Moons, "Clinical prediction models: Diagnosis versus prognosis," *Journal of Clinical Epidemiology*, vol. 132, pp. 142–145, 2021, ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2021.01.009.

[14] M. Kim, C. Rouphael, J. Mcmichael, N. Welch, and S. Dasarathy, "Challenges in and opportunities for electronic health record-based data analysis and interpretation," *Gut and Liver*, vol. 18, pp. 201–208, Oct. 2023. DOI: 10.5009/gnl230272.

[15] A. Subbaswamy and S. Saria, "From development to deployment: Dataset shift, causality, and shift-stable models in health ai," *Biostatistics (Oxford, England)*, vol. 21, Nov. 2019. DOI: 10.1093/biostatistics/kxz041.

[16] D. Agniel, I. Kohane, and G. Weber, "Biases in electronic health record data due to processes within the healthcare system: Retrospective observational study," *BMJ*, vol. 361, k1479, Apr. 2018. DOI: 10.1136/bmj.k1479.

[17] Y. Zhao, Z. S.-Y. Wong, and K. L. Tsui, "A framework of rebalancing imbalanced healthcare data for rare events classification: A case of look-alike sound-alike mix-up incident detection," *Journal of Healthcare Engineering*, vol. 2018, no. 1, p. 6 275 435, 2018. DOI: https://doi.org/10.1155/2018/6275435. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2018/6275435. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2018/6275435.

[18] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, pp. 447–453, Oct. 2019. DOI: 10.1126/science.aax2342.

[19] B. Bischl et al., *Openml benchmarking suites*, 2021. arXiv: 1708.03731 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1708.03731.

[20] P. Gijsbers et al., *Amlb: An automl benchmark*, 2023. arXiv: 2207.12560 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2207.12560.

[21] N. Erickson et al., *Tabarena: A living benchmark for machine learning on tabular data*, 2025. arXiv: 2506.16791 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2506.16791.

[22] V. Borisov, T. Leemann, K. SeSSler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep neural networks and tabular data: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 6, pp. 7499–7519, 2024, ISSN: 2162-2388. DOI: 10.1109/tnnls.2022.3229161. [Online]. Available: http://dx.doi.org/10.1109/TNNLS.2022.3229161.

[23] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Aug. 2016, pp. 785–794. DOI: 10.

1145/2939672.2939785. [Online]. Available: http://dx.doi.org/10.1145/2939672.2939785.

[24] S. Müller, N. Hollmann, S. P. Arango, J. Grabocka, and F. Hutter, *Transformers can do bayesian inference*, 2024. arXiv: 2112.10510 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2112.10510.

[25] A. Garg, M. Ali, N. Hollmann, L. Purucker, S. Müller, and F. Hutter, "Realtabpfn: Improving tabular foundation models via continued pre-training with real-world data," 2025. arXiv: 2507.03971 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2507.03971.

[26] J. Qu, D. Holzmüller, G. Varoquaux, and M. L. Morvan, *Tabicl: A tabular foundation model for in-context learning on large data*, 2025. arXiv: 2502.05564 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2502.05564.

[27] X. Zhang et al., *Mitra: Mixed synthetic priors for enhancing tabular foundation models*, 2025. arXiv: 2510.21204 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2510.21204.

[28] M. Grassi et al., "A novel ensemble-based machine learning algorithm to predict the conversion from mild cognitive impairment to Alzheimer's disease using socio-demographic characteristics, clinical information, and neuropsychological measures," *Frontiers in Neurology*, vol. 10, p. 756, 2019. DOI: 10.3389/fneur.2019.00756.

[29] R. Marinescu et al., "Tadpole challenge: Accurate alzheimers disease prediction through crowdsourced forecasting of future data," vol. 11843, Oct. 2019, pp. 1–10, ISBN: 978-3-030-32280-9. DOI: 10.1007/978-3-030-32281-6_1.

[30] L. Wang, S. Zhang, Z. Gao, and D. Jiang, "Construction and validation of a risk prediction model for chronic obstructive pulmonary disease (copd): A cross-sectional study based on the nhanes database from 2009 to 2018," *BMC Pulmonary Medicine*, vol. 25, Jul. 2025. DOI: 10.1186/s12890-025-03776-w.

[31] T. Vu et al., "Prediction of depressive disorder using machine learning approaches: Findings from the nhanes," *BMC Medical Informatics and Decision Making*, vol. 25, Feb. 2025. DOI: 10.1186/s12911-025-02903-1.

[32] F. Lopez, E. Núñez Valdez, R. Gonzalez Crespo, and V. García Díaz, "An artificial neural network approach for predicting hypertension using nhanes data," *Scientific Reports*, vol. IP, pp. 1–36, Jun. 2020. DOI: 10.1038/s41598-020-67640-z.

[33] M. Ezz, M. Abdullah, M. Alshammeri, A. Tantawy, A. Allahim, and A. Mostafa, "Machine learning framework for hba1c prediction: Data enrichment, cost optimization, and interpretability through stratified regression and multi-stage feature selection," *Diagnostics*, vol. 16, p. 607, Feb. 2026. DOI: 10.3390/diagnostics16040607.

[34] Y. Zhang, X. Liu, X. Zhang, Y. Fei, and X. Li, "Machine learning-based prediction of metabolic dysfunction-associated steatotic liver disease using national health and nutrition examination survey (nhanes) data," *PLOS One*, vol. 20, Nov. 2025. DOI: 10.1371/journal.pone.0335656.

[35] C. Zhang, B. Niu, R. Wang, and L. Zhang, "From traditional metabolic markers to ensemble learning: Comparative application of machine learning models

for predicting nafld risk in adolescents," *Frontiers in Endocrinology*, vol. 16, Oct. 2025. DOI: 10.3389/fendo.2025.1681686.

[36] Y. Fu, Y. Yu, and W. Chen, "Constructing machine learning-based risk prediction model for osteoarthritis in population aged 45 and above: Nhanes 20112018," *Scientific Reports*, vol. 15, Apr. 2025. DOI: 10.1038/s41598-025-99411-z.

[37] A. Johnson et al., "Mimic-iv, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, p. 1, Jan. 2023. DOI: 10.1038/s41597-022-01899-x.

[38] H. Bui, H. Warrier, and Y. Gupta, *Benchmarking with mimic-iv, an irregular, spare clinical time series dataset*, 2024. arXiv: 2401.15290 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2401.15290.

[39] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, "MIMIC-IV," *PhysioNet*, Jan. 2023, Version 2.2. DOI: 10.13026/6mm1-ek67. [Online]. Available: https://doi.org/10.13026/6mm1-ek67.

[40] Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS), *National health and nutrition examination survey data*, Accessed: 2026-04-15, Hyattsville, MD. [Online]. Available: https://wwwn.cdc.gov/nchs/nhanes/Default.aspx.

[41] L. Wang, S. Zhang, Z. Gao, and D. Jiang, "Construction and validation of a risk prediction model for chronic obstructive pulmonary disease (copd): A cross-sectional study based on the nhanes database from 2009 to 2018," *BMC Pulmonary Medicine*, vol. 25, Jul. 2025. DOI: 10.1186/s12890-025-03776-w.

[42] World Health Organization, *Chronic obstructive pulmonary disease (COPD)*, Accessed: 2026-05-17, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd).

[43] World Health Organization, *Depression*, Accessed: 2026-05-17, 2024. [Online]. Available: https://www.who.int/health-topics/depression.

[44] American Diabetes Association, *Diagnosis*, Accessed: 2026-05-18, 2025. [Online]. Available: https://diabetes.org/about-diabetes/diagnosis.

[45] World Health Organization, *Diabetes*, Accessed: 2026-05-18, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes.

[46] World Health Organization, "Use of glycated haemoglobin (HbA1c) in the diagnosis of diabetes mellitus," World Health Organization, Tech. Rep. WHO/NMH/CHP/CPM/2011. [Online]. Available: https://www.who.int/publications/i/item/use-of-glycated-haemoglobin-(-hba1c)-in-diagnosis-of-diabetes-mellitus.

[47] World Health Organization, *Hypertension*, Accessed: 2026-05-17, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hypertension.

[48] M. Rinella et al., "A multi-society delphi consensus statement on new fatty liver disease nomenclature," *Hepatology (Baltimore, Md.)*, vol. Publish Ahead of Print, Jun. 2023. DOI: 10.1097/HEP.0000000000000520.

[49] M. Teng et al., "Global incidence and prevalence of nonalcoholic fatty liver disease," *Clinical and Molecular Hepatology*, vol. 29, S32–S42, Feb. 2023. DOI: 10.3350/cmh.2022.0365.

[50] C. Jack et al., "The alzheimer's disease neuroimaging initiative (adni): Mri methods," *Journal of magnetic resonance imaging : JMRI*, vol. 27, pp. 685–91, May 2008. DOI: 10.1002/jmri.21049.

[51] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, *Optuna: A next-generation hyperparameter optimization framework*, 2019. arXiv: 1907.10902 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1907.10902.

[52] R. Petersen et al., "Alzheimer's disease neuroimaging initiative (adni): Clinical characterization," *Neurology*, vol. 74, pp. 201–9, Jan. 2010. DOI: 10.1212/WNL.0b013e3181cb3e25.

[53] Alzheimer's Disease Neuroimaging Initiative, *Alzheimer's Disease Neuroimaging Initiative Protocol*, https://adni.loni.usc.edu/wp-content/themes/freshnews-dev-v2/documents/clinical/ADNI-1_Protocol.pdf, ADNI protocol. Accessed 18 May 2026, 2005.

[54] Alzheimer's Disease Neuroimaging Initiative, *Inclusion Criteria — ADNIMERGE Documentation*, https://adni.bitbucket.io/reference/inclusio.html, Accessed 18 May 2026, n.d.

[55] National Center for Health Statistics, *NHANES 2017–2018 Laboratory Data: Cotinine and Hydroxycotinine*, https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/COT_J.htm, Accessed 23 May 2026, 2020.

[56] National Center for Health Statistics, *Secondhand Smoke Exposure Higher for Adults Living in Poverty, 2015–2018*, https://www.cdc.gov/nchs/data/databriefs/db396-H.pdf, NCHS Data Brief No. 396. Accessed 23 May 2026, 2021.

[57] National Kidney Foundation, *CKD-EPI Creatinine Equation (2021)*, https://www.kidney.org/ckd-epi-creatinine-equation-2021-0, Accessed 24 May 2026, 2021.

[58] R. Shwartz-Ziv and A. Armon, *Tabular data: Deep learning is not all you need*, 2021. arXiv: 2106.03253 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2106.03253.

A

Appendix 1: Task Catalogue

A.1 Alzheimer's Conversion

A.1.1 Task summary

This task predicts observed conversion from Mild Cognitive Impairment (MCI) to dementia in ADNI participants. The cohort consists of participants diagnosed with EMCI or LMCI at baseline, with a baseline MMSE score of at least 20 and at least 36 months of available follow-up.

A.1.2 Prediction target

The prediction target is a binary converter label. A participant is labeled as a converter if they have any post-baseline follow-up visit with a diagnosis of dementia and an MMSE score between 20 and 26 inclusive. Participants without such an observed post-baseline visit are labeled as non-converters.

The original paper describes converters as participants satisfying probable AD criteria during follow-up and having an MMSE score "between 20 and 2". In this implementation, the MMSE criterion was changed to 20-26 inclusive. This choice follows ADNI documentation and related ADNI publications, where AD/probable AD participants are described as having MMSE scores between 20 and 26 inclusive [52], [53], [54].

Unlike the original study definition, the conversion event is not restricted to occur within 36 months from baseline. The 36-month criterion is used only as a minimum follow-up requirement for cohort inclusion. Therefore, the task should be interpreted as prediction of observed conversion during available follow-up, rather than strict 3-year conversion prediction.

A.1.3 Input table and unit of observation

The input data is derived from the ADNI ADNIMERGE table. Each row in the final task table corresponds to one participant, represented by their baseline visit. Only the first baseline record for each participant is retained.

A.1.4 Cohort definition

Participants are included if they satisfy all of the following criteria:

- a recorded baseline visit.
- at least one visit at or beyond 36 months from baseline.
- baseline diagnosis of EMCI, or LMCI.
- baseline MMSE score of at least 20.
- baseline visit code equal to `bl`.

Participants with missing converter labels after applying the above definition are excluded.

A.1.5 Features

The feature table contains baseline demographic and clinical/neuropsychological variables. The categorical variables are:

- sex.
- baseline diagnostic subtype.
- marital status.

The continuous variables are age, years of education, CDR-SB, ADAS11, ADAS13, ADASQ4, MMSE, RAVLT immediate recall, RAVLT learning, RAVLT forgetting, RAVLT percent forgetting, logical memory delayed recall, digit symbol score, Trail Making Test B score, and FAQ.

Table A.1: Summary statistics (ADConv).

Feature	Level	Type	Overall
ADAS11		continuous	9.86 ± 4.37. missing 2 (0.3%)
ADAS13		continuous	15.89 ± 6.67. missing 3 (0.4%)
ADASQ4		continuous	5.33 ± 2.54
AGE		continuous	72.73 ± 7.48. missing 1 (0.1%)
CDRSB		continuous	1.46 ± 0.90
DIGITSCOR		continuous	38.23 ± 11.21. missing 426 (61.1%)
DX_bl	LMCI	categorical	423 (60.7%)
	EMCI		274 (39.3%)
FAQ		continuous	3.00 ± 4.15. missing 8 (1.1%)
LDELTOTAL		continuous	6.02 ± 3.43
MMSE		continuous	27.71 ± 1.76
PTEDUCAT		continuous	16.04 ± 2.74
PTGENDER	Male	categorical	418 (60.0%)
	Female		279 (40.0%)
PTMARRY	Married	categorical	548 (78.6%)
	Widowed		73 (10.5%)
	Divorced		59 (8.5%)

Continued on next page

Table A.1: Summary statistics (ADConv).

Feature	Level	Type	Overall
	Never married		12 (1.7%)
	Unknown		5 (0.7%)
RAVLT_forgetting		continuous	4.59 ± 2.43. missing 1 (0.1%)
RAVLT_immediate		continuous	35.23 ± 10.66
RAVLT_learning		continuous	4.21 ± 2.56
RAVLT_perc_forgetting		continuous	58.45 ± 32.37. missing 1 (0.1%)
TRABSCOR		continuous	108.77 ± 61.58. missing 6 (0.9%)
converter	False	categorical	441 (63.3%)
	True		256 (36.7%)

A.1.6 Task-specific notes

This task is based on the ADNI MCI-to-AD conversion setup, but differs from the original 3-year conversion definition by allowing conversion at any available post-baseline follow-up visit. This modification was made to retain a sufficient number of positive cases. Consequently, non-converters should be interpreted as participants not observed to satisfy the conversion criterion during available follow-up, not necessarily as participants who never converted.

A.2 ADNI 12-Month Longitudinal Prediction Tasks

A.2.1 Task summary

This group of tasks is inspired by the ADNI TADPOLE challenge. The tasks use information available at an index visit to predict clinical, cognitive, or anatomical outcomes 12 months later. The three prediction targets are diagnosis, ADAS13 score, and ventricular volume.

Each sample is defined by a participant, an index visit, and a 12-month follow-up target. Since multiple samples may originate from the same participant, train-test splitting is performed at the participant level to prevent information leakage across splits.

A.2.2 Relation to TADPOLE

The original TADPOLE challenge considered prediction of future diagnosis, ADAS13, and ventricular volume over a longer forecasting horizon and allowed the use of a large number of ADNI variables, including imaging, biomarker, and laboratory-derived measurements. In this benchmark, we construct simplified TADPOLE-inspired tasks using the same broad target categories but a reduced feature set and shorter prediction horizon.

Only variables available in the `ADNIMERGE` table are used as predictors. This restriction was chosen to reduce sparsity and to make the tasks more comparable to the other tabular benchmark tasks. Many of the additional variables available in the

original TADPOLE setting have substantial missingness, which makes them difficult to use consistently across tabular models without extensive task-specific imputation or feature-selection decisions.

The prediction horizon was reduced to 12 months after the index visit. This choice makes the tasks less dependent on long-range follow-up availability and increases the number of usable participant-visits. In addition, later ADNI visits were included where the required predictor and target information was available.

A.2.3 Shared cohort and unit of observation

The input data is derived from the ADNI `ADNIMERGE` table. Each sample corresponds to a participant at an index visit, identified by the participant identifier `RID` and visit month `Month`. Predictors are taken from the index visit, and the prediction targets are obtained from the corresponding visit 12 months later.

Samples are constructed by joining each participant-visit row to the same participant's row 12 months later. Thus, a row at month t is retained if a corresponding row at month $t + 12$ is available. The resulting unit of observation is a participant-visit pair rather than a participant. Since multiple samples may originate from the same participant, train-test splitting is performed at the participant level.

A.2.4 Shared features

The predictor set consists of the remaining `ADNIMERGE` variables after missingness-based filtering and removal of baseline-suffixed, administrative, identifier, protocol, date, site, and acquisition-related variables. The same shared feature table is used for the three 12-month prediction tasks, with task-specific target variables excluded from the predictors to avoid leakage.

A.2.5 Prediction targets

A.2.5.1 Diagnosis at 12 months

The diagnosis task predicts the diagnostic category recorded 12 months after the index visit. The target variable is `DX_t12`, and the task is formulated as a multiclass classification problem.

A.2.5.2 ADAS13 at 12 months

The ADAS13 task predicts the ADAS13 score recorded 12 months after the index visit. The target variable is `ADAS13_t12`, and the task is formulated as a regression problem.

A.2.5.3 Ventricular volume at 12 months

The ventricular-volume task predicts the ventricular volume measurement recorded 12 months after the index visit. The target variable is `Ventricles_t12`, and the task is formulated as a regression problem.

A.2.6 Task-specific summary statistics

A.2.6.1 Diagnosis prediction

Table A.2: Summary statistics (Ventricles).

Feature	Level	Type	Overall
ADAS11		continuous	10.51 ± 7.21. missing 496 (7.9%)
ADAS13		continuous	16.50 ± 10.25. missing 540 (8.6%)
ADASQ4		continuous	5.15 ± 3.01. missing 486 (7.7%)
AGE		continuous	73.28 ± 7.01. missing 3 (0.0%)
APOE4	0	categorical	3471 (55.2%)
	1		2192 (34.8%)
	2		560 (8.9%)
	Missing		67 (1.1%)
CDRSB		continuous	1.84 ± 2.24. missing 496 (7.9%)
DX	MCI	categorical	3082 (49.0%)
	CN		1641 (26.1%)
	Dementia		1093 (17.4%)
	Missing		474 (7.5%)
DX_t12	MCI	categorical	2763 (43.9%)
	CN		2001 (31.8%)
	Dementia		1526 (24.3%)
EcogPtDivatt		continuous	1.88 ± 0.79. missing 2910 (46.3%)
EcogPtLang		continuous	1.78 ± 0.65. missing 2881 (45.8%)
EcogPtMem		continuous	2.14 ± 0.75. missing 2879 (45.8%)
EcogPtOrgan		continuous	1.57 ± 0.66. missing 2959 (47.0%)
EcogPtPlan		continuous	1.45 ± 0.59. missing 2886 (45.9%)
EcogPtTotal		continuous	1.73 ± 0.57. missing 2879 (45.8%)
EcogPtVisspat		continuous	1.42 ± 0.57. missing 2914 (46.3%)
EcogSPDivatt		continuous	1.97 ± 0.95. missing 2976 (47.3%)
EcogSPLang		continuous	1.69 ± 0.78. missing 2884 (45.9%)
EcogSPMem		continuous	2.13 ± 0.95. missing 2885 (45.9%)
EcogSPOrgan		continuous	1.74 ± 0.91. missing 3036 (48.3%)
EcogSPPlan		continuous	1.66 ± 0.85. missing 2927 (46.5%)
EcogSPTotal		continuous	1.80 ± 0.79. missing 2888 (45.9%)
EcogSPVisspat		continuous	1.55 ± 0.78. missing 2955 (47.0%)
Entorhinal		continuous	3509.95 ± 836.87. missing 2128 (33.8%)
FAQ		continuous	4.68 ± 6.87. missing 496 (7.9%)
Fusiform		continuous	17251.99 ± 2766.07. missing 2128 (33.8%)
Hippocampus		continuous	6720.32 ± 1221.54. missing 1958 (31.1%)
ICV		continuous	1532935.39 ± 169206.62. missing 1274 (20.3%)
LDELTOTAL		continuous	8.30 ± 6.00. missing 851 (13.5%)
MMSE		continuous	27.02 ± 3.37. missing 488 (7.8%)
MOCA		continuous	23.25 ± 4.36. missing 2923 (46.5%)
MidTemp		continuous	19357.92 ± 3033.11. missing 2128 (33.8%)
PTGENDER	Male	categorical	3547 (56.4%)
	Female		2743 (43.6%)
RAVLT_forgetting		continuous	4.27 ± 2.66. missing 528 (8.4%)
RAVLT_immediate		continuous	35.43 ± 13.07. missing 520 (8.3%)
RAVLT_learning		continuous	4.20 ± 2.80. missing 520 (8.3%)
RAVLT_perc_forgetting		continuous	57.53 ± 45.75. missing 554 (8.8%)
TRABSCOR		continuous	115.82 ± 74.37. missing 644 (10.2%)

Continued on next page

Table A.2: Summary statistics (Ventricles).

Feature	Level	Type	Overall
Ventricles		continuous	41978.94 ± 22799.48. missing 1527 (24.3%)
WholeBrain		continuous	1013235.43 ± 110352.64. missing 1404 (22.3%)
mPACCdigit		continuous	-5.72 ± 7.12. missing 483 (7.7%)
mPACCtrailsB		continuous	-5.39 ± 6.77. missing 482 (7.7%)

A.2.6.2 ADAS13 prediction

Table A.3: Summary statistics (ADAS13).

Feature	Level	Type	Overall
ADAS11		continuous	10.35 ± 6.91. missing 490 (7.9%)
ADAS13		continuous	16.31 ± 9.94. missing 528 (8.5%)
ADAS13_t12		continuous	17.45 ± 12.31. missing 0 (0.0%)
ADASQ4		continuous	5.13 ± 2.99. missing 481 (7.8%)
AGE		continuous	73.29 ± 7.00. missing 4 (0.1%)
APOE4	0	categorical	3419 (55.3%)
	1		2155 (34.8%)
	2		540 (8.7%)
	Missing		70 (1.1%)
CDRSB		continuous	1.79 ± 2.14. missing 489 (7.9%)
DX	MCI	categorical	3054 (49.4%)
	CN		1618 (26.2%)
	Dementia		1041 (16.8%)
	Missing		471 (7.6%)
EcogPtDivatt		continuous	1.88 ± 0.79. missing 2845 (46.0%)
EcogPtLang		continuous	1.78 ± 0.65. missing 2816 (45.5%)
EcogPtMem		continuous	2.14 ± 0.75. missing 2814 (45.5%)
EcogPtOrgan		continuous	1.57 ± 0.66. missing 2895 (46.8%)
EcogPtPlan		continuous	1.45 ± 0.59. missing 2822 (45.6%)
EcogPtTotal		continuous	1.73 ± 0.57. missing 2815 (45.5%)
EcogPtVisspat		continuous	1.42 ± 0.57. missing 2850 (46.1%)
EcogSPDivatt		continuous	1.96 ± 0.95. missing 2909 (47.0%)
EcogSPLang		continuous	1.69 ± 0.77. missing 2820 (45.6%)
EcogSPMem		continuous	2.12 ± 0.95. missing 2823 (45.7%)
EcogSPOrgan		continuous	1.73 ± 0.91. missing 2972 (48.1%)
EcogSPPlan		continuous	1.65 ± 0.85. missing 2863 (46.3%)
EcogSPTotal		continuous	1.79 ± 0.78. missing 2824 (45.7%)
EcogSPVisspat		continuous	1.54 ± 0.77. missing 2889 (46.7%)
Entorhinal		continuous	3514.23 ± 836.15. missing 2090 (33.8%)
FAQ		continuous	4.54 ± 6.70. missing 488 (7.9%)
Fusiform		continuous	17269.24 ± 2752.80. missing 2090 (33.8%)
Hippocampus		continuous	6726.71 ± 1221.72. missing 1923 (31.1%)
ICV		continuous	1532348.92 ± 168963.55. missing 1253 (20.3%)
LDELTOTAL		continuous	8.34 ± 5.98. missing 841 (13.6%)
MMSE		continuous	27.10 ± 3.21. missing 483 (7.8%)
MOCA		continuous	23.31 ± 4.24. missing 2863 (46.3%)
MidTemp		continuous	19377.71 ± 3023.44. missing 2090 (33.8%)
PTGENDER	Male	categorical	3490 (56.4%)

Continued on next page

Table A.3: Summary statistics (ADAS13).

Feature	Level	Type	Overall
	Female		2694 (43.6%)
RAVLT_forgetting		continuous	4.30 ± 2.64. missing 519 (8.4%)
RAVLT_immediate		continuous	35.58 ± 12.89. missing 512 (8.3%)
RAVLT_learning		continuous	4.22 ± 2.79. missing 512 (8.3%)
RAVLT_perc_forgetting		continuous	57.54 ± 40.31. missing 538 (8.7%)
TRABSCOR		continuous	115.12 ± 73.65. missing 617 (10.0%)
Ventricles		continuous	41789.73 ± 22680.93. missing 1505 (24.3%)
WholeBrain		continuous	1013651.94 ± 110136.00. missing 1381 (22.3%)
mPACCdigit		continuous	-5.59 ± 6.92. missing 478 (7.7%)
mPACCtrailsB		continuous	-5.26 ± 6.56. missing 477 (7.7%)

A.2.6.3 Ventricles prediction

Table A.4: Summary statistics (DX).

Feature	Level	Type	Overall
ADAS11		continuous	10.10 ± 6.68. missing 296 (6.7%)
ADAS13		continuous	15.93 ± 9.69. missing 325 (7.3%)
ADASQ4		continuous	5.05 ± 2.99. missing 289 (6.5%)
AGE		continuous	73.23 ± 6.97. missing 4 (0.1%)
APOE4	0	categorical	2435 (55.0%)
	1		1569 (35.4%)
	2		395 (8.9%)
	Missing		28 (0.6%)
CDRSB		continuous	1.64 ± 1.95. missing 290 (6.6%)
DX	MCI	categorical	2172 (49.1%)
	CN		1286 (29.0%)
	Dementia		685 (15.5%)
	Missing		284 (6.4%)
EcogPtDivatt		continuous	1.86 ± 0.78. missing 2334 (52.7%)
EcogPtLang		continuous	1.77 ± 0.64. missing 2323 (52.5%)
EcogPtMem		continuous	2.11 ± 0.74. missing 2321 (52.4%)
EcogPtOrgan		continuous	1.54 ± 0.63. missing 2360 (53.3%)
EcogPtPlan		continuous	1.43 ± 0.56. missing 2323 (52.5%)
EcogPtTotal		continuous	1.71 ± 0.55. missing 2320 (52.4%)
EcogPtVisspat		continuous	1.40 ± 0.55. missing 2336 (52.8%)
EcogSPDivatt		continuous	1.88 ± 0.90. missing 2364 (53.4%)
EcogSPLang		continuous	1.62 ± 0.72. missing 2324 (52.5%)
EcogSPMem		continuous	2.03 ± 0.91. missing 2324 (52.5%)
EcogSPOrgan		continuous	1.64 ± 0.83. missing 2411 (54.5%)
EcogSPPlan		continuous	1.57 ± 0.78. missing 2347 (53.0%)
EcogSPTotal		continuous	1.71 ± 0.72. missing 2325 (52.5%)
EcogSPVisspat		continuous	1.47 ± 0.70. missing 2360 (53.3%)
Entorhinal		continuous	3521.09 ± 836.61. missing 1149 (26.0%)
FAQ		continuous	4.14 ± 6.24. missing 290 (6.6%)
Fusiform		continuous	17229.61 ± 2733.92. missing 1149 (26.0%)
Hippocampus		continuous	6754.23 ± 1203.77. missing 1095 (24.7%)
ICV		continuous	1532566.58 ± 168372.65. missing 568 (12.8%)

Continued on next page

Table A.4: Summary statistics (DX).

Feature	Level	Type	Overall
LDELTOTAL		continuous	8.34 ± 5.96. missing 592 (13.4%)
MMSE		continuous	27.20 ± 3.10. missing 292 (6.6%)
MOCA		continuous	23.65 ± 4.05. missing 2353 (53.2%)
MidTemp		continuous	19360.09 ± 2987.57. missing 1149 (26.0%)
PTGENDER	Male	categorical	2484 (56.1%)
	Female		1943 (43.9%)
RAVLT_forgetting		continuous	4.25 ± 2.62. missing 309 (7.0%)
RAVLT_immediate		continuous	35.83 ± 12.90. missing 305 (6.9%)
RAVLT_learning		continuous	4.30 ± 2.83. missing 305 (6.9%)
RAVLT_perc_forgetting		continuous	56.44 ± 47.58. missing 324 (7.3%)
TRABSCOR		continuous	113.33 ± 71.73. missing 375 (8.5%)
Ventricles		continuous	40719.18 ± 21733.72. missing 683 (15.4%)
Ventricles_t12		continuous	43426.74 ± 23407.15. missing 0 (0.0%)
WholeBrain		continuous	1013368.44 ± 108213.36. missing 626 (14.1%)
mPACCdigit		continuous	-5.31 ± 6.66. missing 287 (6.5%)
mPACCtrailsB		continuous	-5.08 ± 6.40. missing 287 (6.5%)

A.2.7 Task-specific notes

These tasks should be interpreted as simplified TADPOLE-inspired prediction tasks rather than reproductions of the original challenge. They use the same broad target variables as TADPOLE but differ in prediction horizon, feature set, and sample construction. The fixed 12-month horizon and `ADNIMERGE`-based feature set were chosen to reduce sparsity, increase reproducibility, and make the tasks suitable for systematic comparison of tabular prediction models.

A.3 YAIB-adopted MIMIC-IV tasks

A.3.1 Task summary

The MIMIC-IV tasks are derived from ICU stays and use longitudinal clinical measurements to predict clinical outcomes. Unlike the ADNI tasks, where each row is already a visit-level tabular observation, the MIMIC-IV tasks are constructed from irregularly sampled time-series data. These time-series measurements are first mapped to an hourly grid and then exported as dynamic and static cohort tables.

Each sample is defined by an ICU stay and a time point. Static variables describe patient- or stay-level information, while dynamic variables describe time-varying clinical measurements. Train-test splitting is performed at the stay level to prevent observations from the same ICU stay appearing in both training and test data.

A.3.2 Shared cohort and unit of observation

The base cohort is constructed from MIMIC-IV ICU stay windows using the `ricu` package. Stay windows are discretized into one-hour intervals and truncated to a

maximum duration of seven days. The resulting dynamic table contains one row per stay-hour pair, while the static table contains one row per ICU stay.

The base cohort is used to define the set of eligible ICU stays shared by the MIMIC-IV tasks. Task-specific cohorts are then derived by restricting to stays that remain after the base-cohort exclusions and by applying any additional task-specific exclusion criteria.

The base cohort applies the following exclusion criteria:

- invalid ICU length of stay.
- ICU stay shorter than 6 hours.
- fewer than four observed dynamic measurements.
- gaps longer than 12 hours without dynamic measurements.
- age below 18 years.

These exclusions remove stays with invalid or insufficient observation history before task-specific outcomes are constructed. Attrition counts are stored separately for the base cohort.

A.3.3 Shared predictors

The MIMIC-IV tasks use the same static and dynamic predictors. Static predictors in the base cohort include age, sex, ethnicity, admission information, ICU length of stay, and hospital length of stay. Dynamic predictors are loaded from the shared dynamic variable dictionary and represent time-varying clinical measurements.

Dynamic variables are mapped to a regular hourly grid. This produces a common time axis across ICU stays and allows irregularly sampled measurements to be represented in a fixed sequential format.

A.3.4 Preprocessing and tabularization

The hourly MIMIC-IV time-series data was converted into fixed-length tabular feature vectors following the YAIB tabularization scheme. Each dynamic variable was represented by summary features describing its last observed value, missingness indicators, maximum, minimum, and mean over the observation window. Static variables were joined at the stay level.

For dynamic outcome tasks, such as sepsis, AKI, and length of stay, labels are defined at the stay-hour level. For static outcome tasks (kidney function and mortality prediction), the dynamic predictors are summarized over the observation window and only the final sample is retained per ICU stay.

A.3.5 Sepsis prediction

A.3.5.1 Task summary

The sepsis task predicts future sepsis onset during an ICU stay. It is a dynamic binary prediction task: both predictors and outcomes are indexed over time, and the output is a label for each stay-hour observation.

A.3.5.2 Prediction target

The sepsis outcome is derived from the `sep3_alt` concept. For each ICU stay, the first sepsis occurrence is mapped to the hourly grid. At each observation time, the binary label indicates whether sepsis onset occurs within the following 6 hours.

A.3.5.3 Additional exclusions

In addition to the base-cohort exclusions, the sepsis task excludes stays with sepsis onset before 6 hours in the ICU.

A.3.5.4 Summary statistics

Table A.5: Summary statistics (Sepsis).

Feature	Level	Type	Overall
alb		continuous	3.15 ± 0.65; missing 2413842 (99.2%)
alp		continuous	120.22 ± 130.02; missing 2397607 (98.5%)
alt		continuous	244.60 ± 929.28; missing 2396974 (98.5%)
ast		continuous	298.54 ± 1253.25; missing 2396701 (98.5%)
be		continuous	-0.61 ± 5.19; missing 2336729 (96.0%)
bicar		continuous	24.05 ± 4.84; missing 2272417 (93.4%)
bili		continuous	2.32 ± 5.26; missing 2397088 (98.5%)
bili_dir		continuous	3.28 ± 4.75; missing 2431335 (99.9%)
bnd		continuous	3.93 ± 6.58; missing 2428912 (99.8%)
bun		continuous	26.27 ± 22.09; missing 2271533 (93.3%)
ca		continuous	8.43 ± 0.77; missing 2287259 (94.0%)
cai		continuous	1.13 ± 0.11; missing 2373473 (97.5%)
ck		continuous	1855.87 ± 9100.89; missing 2410049 (99.0%)
ckmb		continuous	23.68 ± 55.27; missing 2409517 (99.0%)
cl		continuous	103.18 ± 6.60; missing 2265307 (93.1%)
crea		continuous	1.40 ± 1.45; missing 2271041 (93.3%)
crp		continuous	69.10 ± 79.04; missing 2432263 (99.9%)
dbp		continuous	64.84 ± 15.34; missing 349400 (14.4%)
fgn		continuous	279.58 ± 159.70; missing 2418537 (99.4%)
fio2		continuous	50.62 ± 18.78; missing 2245688 (92.3%)
glu		continuous	139.34 ± 58.97; missing 2239884 (92.0%)

Continued on next page

Table A.5: Summary statistics (Sepsis).

Feature	Level	Type	Overall
hgb		continuous	10.25 ± 2.09; missing 2280364 (93.7%)
hr		continuous	83.58 ± 17.87; missing 192782 (7.9%)
inr_pt		continuous	1.47 ± 0.80; missing 2328994 (95.7%)
k		continuous	4.13 ± 0.64; missing 2261382 (92.9%)
outcome	False	categorical	2337844 (96.1%)
	True		95884 (3.9%)
lact		continuous	2.42 ± 2.24; missing 2364918 (97.2%)
lymph		continuous	14.18 ± 12.01; missing 2413307 (99.2%)
map		continuous	80.09 ± 15.54; missing 320936 (13.2%)
mch		continuous	29.96 ± 2.54; missing 2283254 (93.8%)
mchc		continuous	32.99 ± 1.64; missing 2283214 (93.8%)
mcv		continuous	90.86 ± 6.76; missing 2283238 (93.8%)
methb		continuous	2.30 ± 5.00; missing 2433472 (100.0%)
mg		continuous	2.07 ± 0.36; missing 2277889 (93.6%)
na		continuous	138.34 ± 5.58; missing 2262781 (93.0%)
neut		continuous	76.26 ± 15.49; missing 2413307 (99.2%)
o2sat		continuous	95.99 ± 3.39; missing 245193 (10.1%)
pco2		continuous	42.24 ± 11.45; missing 2336644 (96.0%)
ph		continuous	7.38 ± 0.08; missing 2323383 (95.5%)
phos		continuous	3.54 ± 1.35; missing 2286451 (93.9%)
plt		continuous	200.47 ± 107.49; missing 2280591 (93.7%)
po2		continuous	148.47 ± 97.25; missing 2343931 (96.3%)
ptt		continuous	41.96 ± 25.57; missing 2321519 (95.4%)
resp		continuous	19.11 ± 5.41; missing 221157 (9.1%)
sbp		continuous	121.27 ± 21.66; missing 348994 (14.3%)
temp		continuous	36.87 ± 0.59; missing 1790725 (73.6%)
tnt		continuous	1.04 ± 2.34; missing 2414186 (99.2%)
urine		continuous	156.82 ± 168.04; missing 1259564 (51.8%)
wbc		continuous	11.39 ± 9.47; missing 2283038 (93.8%)
age		continuous	62.51 ± 17.37
height		continuous	168.57 ± 13.89; missing 30082 (59.4%)
sex	Male	categorical	27569 (54.4%)
	Female		23106 (45.6%)
weight		continuous	80.30 ± 23.29; missing 3309 (6.5%)

A.3.6 Acute kidney injury prediction

A.3.6.1 Task summary

The acute kidney injury (AKI) task predicts future AKI onset during an ICU stay. Like the sepsis task, it is a dynamic binary prediction task: predictors and outcomes are indexed over time, and the output is a binary label for each stay-hour observation.

A.3.6.2 Prediction target

The AKI outcome is derived from the `aki` concept. For each ICU stay, the first AKI occurrence is mapped to the hourly grid. At each observation time, the binary label indicates whether AKI onset occurs within the following 6 hours.

A.3.6.3 Additional exclusions

In addition to the base-cohort exclusions, the AKI task excludes stays with AKI onset before 6 hours in the ICU and stays with baseline creatinine greater than 4 mg/dL.

A.3.6.4 Summary statistics

Table A.6: Summary statistics (AKI).

Feature	Level	Type	Overall
alb		continuous	3.07 ± 0.65; missing 2681715 (99.1%)
alp		continuous	118.99 ± 127.71; missing 2660984 (98.3%)
alt		continuous	221.60 ± 877.49; missing 2660131 (98.3%)
ast		continuous	267.78 ± 1054.13; missing 2659952 (98.3%)
be		continuous	-0.55 ± 5.07; missing 2548833 (94.2%)
bicar		continuous	24.03 ± 4.88; missing 2517109 (93.0%)
bili		continuous	1.94 ± 3.67; missing 2660341 (98.3%)
bili_dir		continuous	2.79 ± 3.82; missing 2703535 (99.9%)
bnd		continuous	5.09 ± 7.95; missing 2699268 (99.7%)
bun		continuous	22.65 ± 17.37; missing 2516133 (93.0%)
ca		continuous	8.32 ± 0.76; missing 2538622 (93.8%)
cai		continuous	1.13 ± 0.10; missing 2615364 (96.6%)
ck		continuous	1709.30 ± 8795.50; missing 2677048 (98.9%)
ckmb		continuous	23.54 ± 55.05; missing 2677167 (98.9%)
cl		continuous	104.52 ± 6.49; missing 2506119 (92.6%)
crea		continuous	1.04 ± 0.61; missing 2515429 (92.9%)
crp		continuous	79.97 ± 81.82; missing 2705299 (99.9%)
dbp		continuous	63.98 ± 14.68; missing 319399 (11.8%)
fgn		continuous	276.45 ± 161.98; missing 2685512 (99.2%)
fio2		continuous	51.07 ± 18.59; missing 2399280 (88.6%)
glu		continuous	140.41 ± 58.78; missing 2462427 (91.0%)
hgb		continuous	10.25 ± 2.02; missing 2522812 (93.2%)
hr		continuous	84.62 ± 17.85; missing 177352 (6.6%)
inr_pt		continuous	1.47 ± 0.79; missing 2581304 (95.4%)
k		continuous	4.09 ± 0.62; missing 2502380 (92.4%)
outcome	False	categorical	2373184 (87.7%)
	True		333762 (12.3%)
lact		continuous	2.45 ± 2.12; missing 2603283 (96.2%)

Continued on next page

Table A.6: Summary statistics (AKI).

Feature	Level	Type	Overall
lymph		continuous	13.60 ± 12.10; missing 2678188 (98.9%)
map		continuous	79.39 ± 15.06; missing 277953 (10.3%)
mch		continuous	30.03 ± 2.54; missing 2526063 (93.3%)
mchc		continuous	33.07 ± 1.66; missing 2526015 (93.3%)
mcv		continuous	90.86 ± 6.76; missing 2526042 (93.3%)
methb		continuous	1.13 ± 4.24; missing 2706604 (100.0%)
mg		continuous	2.05 ± 0.36; missing 2525470 (93.3%)
na		continuous	138.89 ± 5.61; missing 2504349 (92.5%)
neut		continuous	76.85 ± 15.71; missing 2678189 (98.9%)
o2sat		continuous	96.15 ± 3.34; missing 216094 (8.0%)
pco2		continuous	42.43 ± 11.39; missing 2548715 (94.2%)
ph		continuous	7.38 ± 0.08; missing 2531858 (93.5%)
phos		continuous	3.30 ± 1.13; missing 2537396 (93.7%)
plt		continuous	196.19 ± 107.90; missing 2522573 (93.2%)
po2		continuous	149.79 ± 97.56; missing 2557834 (94.5%)
ptt		continuous	40.47 ± 24.17; missing 2575005 (95.1%)
resp		continuous	19.38 ± 5.57; missing 197785 (7.3%)
sbp		continuous	120.27 ± 21.17; missing 318980 (11.8%)
temp		continuous	36.96 ± 0.67; missing 1954656 (72.2%)
tnt		continuous	0.93 ± 2.28; missing 2682988 (99.1%)
urine		continuous	147.05 ± 158.49; missing 1163754 (43.0%)
wbc		continuous	11.91 ± 9.68; missing 2525789 (93.3%)
age		continuous	63.28 ± 17.03
height		continuous	168.83 ± 13.34; missing 33685 (53.6%)
sex	Male	categorical	34651 (55.2%)
	Female		28170 (44.8%)
weight		continuous	80.62 ± 23.35; missing 3926 (6.2%)

A.3.7 Length of stay prediction

A.3.7.1 Task summary

The length of stay (LOS) regression task predicts remaining ICU length of stay during an ICU admission. Like the sepsis and AKI tasks, it is a dynamic prediction task: predictors and outcomes are indexed over time, and the output is a label for each stay-hour observation.

A.3.7.2 Prediction target

The LOS outcome is derived from the `los_icu` concept. ICU length of stay is converted to hours and mapped to the hourly grid. At each observation time, the label is the remaining ICU length of stay in hours, computed as the total ICU length of stay minus the current time. The target is capped at 7 days, corresponding to

the maximum observation window.

A.3.7.3 Additional exclusions

No task-specific exclusions are applied beyond the base-cohort exclusions.

A.3.7.4 Summary statistics

Table A.7: Summary statistics (LOS).

Feature	Level	Type	Overall
alb		continuous	2.98 ± 0.65; missing 4676363 (99.1%)
alp		continuous	130.07 ± 142.28; missing 4634980 (98.2%)
alt		continuous	268.54 ± 901.92; missing 4633850 (98.2%)
ast		continuous	364.18 ± 1340.97; missing 4633185 (98.2%)
be		continuous	-0.82 ± 5.41; missing 4438775 (94.0%)
bicar		continuous	23.84 ± 5.15; missing 4386705 (92.9%)
bili		continuous	3.07 ± 6.00; missing 4633250 (98.2%)
bili_dir		continuous	3.82 ± 5.02; missing 4712915 (99.8%)
bnd		continuous	4.66 ± 7.48; missing 4706584 (99.7%)
bun		continuous	29.70 ± 23.82; missing 4386325 (92.9%)
ca		continuous	8.35 ± 0.80; missing 4419218 (93.6%)
cai		continuous	1.12 ± 0.11; missing 4554631 (96.5%)
ck		continuous	2557.90 ± 12639.17; missing 4676607 (99.1%)
ckmb		continuous	22.54 ± 52.79; missing 4677651 (99.1%)
cl		continuous	103.60 ± 6.85; missing 4369832 (92.6%)
crea		continuous	1.55 ± 1.49; missing 4385295 (92.9%)
crp		continuous	88.61 ± 83.27; missing 4717994 (99.9%)
dbp		continuous	63.19 ± 14.88; missing 570049 (12.1%)
fgn		continuous	282.06 ± 174.30; missing 4681123 (99.2%)
fio2		continuous	50.33 ± 17.88; missing 4144099 (87.8%)
glu		continuous	140.80 ± 59.37; missing 4306674 (91.2%)
hgb		continuous	9.94 ± 1.97; missing 4410217 (93.4%)
hr		continuous	85.25 ± 17.97; missing 289241 (6.1%)
inr_pt		continuous	1.56 ± 0.88; missing 4507166 (95.5%)
k		continuous	4.13 ± 0.65; missing 4362130 (92.4%)
label		continuous	67.75 ± 58.80
lact		continuous	2.62 ± 2.52; missing 4533776 (96.0%)
lymph		continuous	13.02 ± 12.40; missing 4677567 (99.1%)
map		continuous	78.83 ± 15.31; missing 475201 (10.1%)
mch		continuous	30.02 ± 2.53; missing 4416479 (93.6%)
mchc		continuous	32.97 ± 1.69; missing 4416388 (93.6%)
mcv		continuous	91.13 ± 6.89; missing 4416444 (93.6%)
methb		continuous	1.64 ± 4.31; missing 4719813 (100.0%)

Continued on next page

Table A.7: Summary statistics (LOS).

Feature	Level	Type	Overall
mg		continuous	2.10 ± 0.37; missing 4397507 (93.2%)
na		continuous	138.73 ± 5.77; missing 4365233 (92.5%)
neut		continuous	76.62 ± 16.49; missing 4677568 (99.1%)
o2sat		continuous	96.06 ± 3.48; missing 372693 (7.9%)
pco2		continuous	42.18 ± 11.31; missing 4438524 (94.0%)
ph		continuous	7.37 ± 0.09; missing 4409817 (93.4%)
phos		continuous	3.61 ± 1.48; missing 4417062 (93.6%)
plt		continuous	186.11 ± 110.87; missing 4409760 (93.4%)
po2		continuous	133.60 ± 85.98; missing 4455420 (94.4%)
ptt		continuous	43.22 ± 25.20; missing 4492163 (95.2%)
resp		continuous	19.58 ± 5.71; missing 326322 (6.9%)
sbp		continuous	119.90 ± 21.68; missing 569227 (12.1%)
temp		continuous	36.94 ± 0.67; missing 3406892 (72.2%)
tnt		continuous	0.95 ± 2.27; missing 4683779 (99.2%)
urine		continuous	139.21 ± 152.53; missing 2147880 (45.5%)
wbc		continuous	12.15 ± 9.76; missing 4415952 (93.6%)
age		continuous	63.21 ± 16.83
height		continuous	168.88 ± 13.27; missing 38529 (53.5%)
sex	Male	categorical	40157 (55.8%)
	Female		31808 (44.2%)
weight		continuous	80.90 ± 23.47; missing 4859 (6.8%)

A.3.8 Kidney function prediction

A.3.8.1 Task summary

The kidney function task predicts post-admission kidney function during an ICU stay. Unlike the sepsis, AKI, and LOS tasks, the outcome is static at the stay level: each ICU stay receives a single regression target. Predictors are still derived from dynamic measurements during the early ICU stay.

A.3.8.2 Prediction target

The outcome is derived from the `crea` concept. For each ICU stay, the target is the median creatinine value observed between 24 and 48 hours after ICU admission. This value is used as a stay-level regression label.

A.3.8.3 Additional exclusions

In addition to the base-cohort exclusions, the kidney function task excludes ICU stays shorter than 48 hours and stays without a creatinine measurement between 24 and 48 hours.

A.3.8.4 Summary statistics

Table A.8: Summary statistics (Kidney Function).

Feature	Level	Type	Overall
alb		continuous	3.03 ± 0.65; missing 835095 (98.4%)
alp		continuous	121.71 ± 134.96; missing 824414 (97.1%)
alt		continuous	288.47 ± 1069.17; missing 824085 (97.1%)
ast		continuous	440.19 ± 1604.86; missing 823945 (97.1%)
be		continuous	-1.70 ± 5.17; missing 752465 (88.7%)
bicar		continuous	22.63 ± 5.04; missing 769657 (90.7%)
bili		continuous	2.54 ± 5.29; missing 824074 (97.1%)
bili_dir		continuous	3.51 ± 4.65; missing 846489 (99.7%)
bnd		continuous	6.20 ± 8.82; missing 844055 (99.4%)
bun		continuous	30.23 ± 24.20; missing 769414 (90.6%)
ca		continuous	8.27 ± 0.87; missing 779545 (91.8%)
cai		continuous	1.13 ± 0.12; missing 796099 (93.8%)
ck		continuous	2509.50 ± 13977.08; missing 831195 (97.9%)
ckmb		continuous	25.32 ± 56.61; missing 830255 (97.8%)
cl		continuous	103.98 ± 6.91; missing 766500 (90.3%)
crea		continuous	1.63 ± 1.58; missing 769169 (90.6%)
crp		continuous	83.04 ± 85.21; missing 847960 (99.9%)
dbp		continuous	62.25 ± 14.75; missing 82221 (9.7%)
fgn		continuous	265.51 ± 156.47; missing 834115 (98.3%)
fio2		continuous	54.11 ± 19.86; missing 705812 (83.2%)
glu		continuous	147.53 ± 65.53; missing 737395 (86.9%)
hgb		continuous	10.11 ± 2.08; missing 770345 (90.8%)
hr		continuous	86.34 ± 18.90; missing 47627 (5.6%)
inr_pt		continuous	1.59 ± 0.92; missing 789390 (93.0%)
k		continuous	4.24 ± 0.73; missing 765563 (90.2%)
lact		continuous	2.79 ± 2.31; missing 778429 (91.7%)
lymph		continuous	12.03 ± 11.19; missing 833469 (98.2%)
map		continuous	77.47 ± 15.03; missing 65561 (7.7%)
mch		continuous	30.04 ± 2.57; missing 771900 (90.9%)
mchc		continuous	33.00 ± 1.73; missing 771870 (90.9%)
mcv		continuous	91.12 ± 7.00; missing 771880 (90.9%)
methb		continuous	0.60 ± 2.89; missing 848552 (100.0%)
mg		continuous	2.08 ± 0.43; missing 775002 (91.3%)
na		continuous	138.26 ± 5.83; missing 766751 (90.3%)
neut		continuous	78.17 ± 15.19; missing 833469 (98.2%)
o2sat		continuous	96.33 ± 3.55; missing 54244 (6.4%)
pco2		continuous	41.96 ± 11.51; missing 752392 (88.6%)
ph		continuous	7.36 ± 0.09; missing 746672 (88.0%)
phos		continuous	3.84 ± 1.55; missing 779168 (91.8%)

Continued on next page

Table A.8: Summary statistics (Kidney Function).

Feature	Level	Type	Overall
plt		continuous	190.16 ± 111.27; missing 769879 (90.7%)
po2		continuous	151.14 ± 98.14; missing 758321 (89.3%)
ptt		continuous	42.08 ± 25.93; missing 787260 (92.8%)
resp		continuous	19.50 ± 5.70; missing 50969 (6.0%)
sbp		continuous	117.07 ± 21.16; missing 82066 (9.7%)
temp		continuous	36.92 ± 0.75; missing 588019 (69.3%)
tnt		continuous	0.96 ± 2.41; missing 832512 (98.1%)
urine		continuous	123.24 ± 147.66; missing 349346 (41.2%)
wbc		continuous	13.19 ± 10.82; missing 771768 (90.9%)
age		continuous	64.09 ± 16.17
height		continuous	168.80 ± 12.62; missing 14203 (41.8%)
outcome		continuous	1.47 ± 1.42
sex	Male	categorical	19147 (56.4%)
	Female		14804 (43.6%)
weight		continuous	81.72 ± 24.24; missing 2709 (8.0%)

A.3.8.5 Task-specific notes

In the final tabular benchmark representation, this task is represented using the final sample per ICU stay, summarizing the entire stay.

A.3.9 Mortality prediction

A.3.9.1 Task summary

The mortality task predicts ICU mortality from information available during the early ICU stay. Like the kidney function task, the outcome is static at the stay level: each ICU stay receives a single binary label. Predictors are derived from dynamic measurements during the first 24 hours of the ICU stay.

A.3.9.2 Prediction target

The outcome is derived from the `death_icu` concept. Each ICU stay is labeled as positive if ICU death is recorded after the prediction window, and negative otherwise. The task is formulated as a stay-level binary classification problem.

A.3.9.3 Additional exclusions

In addition to the base-cohort exclusions, the mortality task excludes stays where death occurred within the first 30 hours of ICU admission and stays with ICU length of stay shorter than 30 hours.

Table A.9: Summary statistics (Mortality).

Feature	Level	Type	Overall
alb		continuous	3.06 ± 0.66; missing 1219914 (98.5%)
alp		continuous	122.11 ± 135.75; missing 1205562 (97.4%)
alt		continuous	265.59 ± 993.96; missing 1205100 (97.3%)
ast		continuous	399.96 ± 1478.90; missing 1204924 (97.3%)
be		continuous	-1.62 ± 5.10; missing 1116560 (90.2%)
bicar		continuous	22.75 ± 4.94; missing 1128227 (91.1%)
bili		continuous	2.44 ± 5.09; missing 1205135 (97.3%)
bili_dir		continuous	3.36 ± 4.45; missing 1235180 (99.8%)
bnd		continuous	5.87 ± 8.61; missing 1232045 (99.5%)
bun		continuous	28.96 ± 23.65; missing 1127880 (91.1%)
ca		continuous	8.29 ± 0.85; missing 1142634 (92.3%)
cai		continuous	1.13 ± 0.12; missing 1170281 (94.5%)
ck		continuous	2204.88 ± 12418.09; missing 1214623 (98.1%)
ckmb		continuous	25.13 ± 57.25; missing 1213324 (98.0%)
cl		continuous	104.02 ± 6.76; missing 1123660 (90.8%)
crea		continuous	1.57 ± 1.56; missing 1127627 (91.1%)
crp		continuous	77.06 ± 83.54; missing 1237017 (99.9%)
dbp		continuous	62.72 ± 14.87; missing 121354 (9.8%)
fgn		continuous	264.64 ± 153.63; missing 1219171 (98.5%)
fio2		continuous	53.98 ± 19.93; missing 1061539 (85.7%)
glu		continuous	145.95 ± 65.07; missing 1085150 (87.6%)
hgb		continuous	10.17 ± 2.08; missing 1128669 (91.2%)
hr		continuous	85.59 ± 18.50; missing 68957 (5.6%)
inr_pt		continuous	1.55 ± 0.89; missing 1156800 (93.4%)
k		continuous	4.23 ± 0.72; missing 1122433 (90.7%)
lact		continuous	2.74 ± 2.32; missing 1150246 (92.9%)
lymph		continuous	12.71 ± 11.49; missing 1217002 (98.3%)
map		continuous	77.79 ± 15.08; missing 100932 (8.2%)
mch		continuous	30.05 ± 2.58; missing 1130754 (91.3%)
mchc		continuous	33.04 ± 1.71; missing 1130714 (91.3%)
mcv		continuous	91.04 ± 6.95; missing 1130735 (91.3%)
methb		continuous	0.94 ± 3.64; missing 1237871 (100.0%)
mg		continuous	2.07 ± 0.42; missing 1136378 (91.8%)
na		continuous	138.20 ± 5.68; missing 1124483 (90.8%)
neut		continuous	77.65 ± 15.30; missing 1217003 (98.3%)
o2sat		continuous	96.29 ± 3.46; missing 80497 (6.5%)
pco2		continuous	41.93 ± 11.24; missing 1116470 (90.2%)
ph		continuous	7.36 ± 0.09; missing 1108023 (89.5%)
phos		continuous	3.77 ± 1.52; missing 1142144 (92.2%)
plt		continuous	190.88 ± 109.30; missing 1128097 (91.1%)
po2		continuous	154.94 ± 100.57; missing 1124129 (90.8%)

Continued on next page

Table A.9: Summary statistics (Mortality.

Feature	Level	Type	Overall
ptt		continuous	41.24 ± 25.34; missing 1154101 (93.2%)
resp		continuous	19.32 ± 5.60; missing 75162 (6.1%)
sbp		continuous	117.72 ± 21.09; missing 121143 (9.8%)
temp		continuous	36.90 ± 0.72; missing 863445 (69.7%)
tnt		continuous	0.93 ± 2.34; missing 1216713 (98.3%)
urine		continuous	130.36 ± 153.18; missing 537188 (43.4%)
wbc		continuous	12.82 ± 10.52; missing 1130579 (91.3%)
age		continuous	63.75 ± 16.45
height		continuous	168.86 ± 12.91; missing 23488 (47.4%)
outcome	0	categorical	45942 (92.8%)
	1		3584 (7.2%)
sex	Male	categorical	27926 (56.4%)
	Female		21600 (43.6%)
weight		continuous	81.35 ± 23.86; missing 3639 (7.3%)

A.3.9.4 Task-specific notes

As in the kidney function task, in the final tabular benchmark representation, this task is represented using the final sample per ICU stay, summarizing the entire stay.

A.4 COPD

A.4.1 Task summary

This task predicts self-reported chronic obstructive pulmonary disease (COPD) in adult NHANES participants. The cohort consists of participants aged 18 years or older from the 2005 through 2018 NHANES survey cycles.

A.4.2 Prediction target

The prediction target is a binary COPD label. A participant is labeled as positive if they report having been told that they had emphysema, chronic bronchitis, or COPD. Participants reporting no to at least one of these questionnaire items and no positive response are labeled as negative. Participants for whom the COPD label cannot be determined are excluded.

The task follows the COPD outcome definition used by Wang et al. [41], with feature construction adapted to the available NHANES files and benchmark preprocessing pipeline.

A.4.3 Input table and unit of observation

The input data is derived from NHANES survey, examination, questionnaire, and laboratory files. Data from multiple survey cycles are combined into a single table. Each row in the final task table corresponds to one NHANES participant, identified by survey cycle and participant identifier.

A.4.4 Cohort definition

Participants are included if they satisfy all of the following criteria:

- participation in one of the included NHANES survey cycles from 2005 to 2018.
- age of at least 18 years.
- sufficient questionnaire information to derive the COPD label.

Participants with missing COPD labels after applying the above definition are excluded.

A.4.5 Features

The feature table contains demographic, questionnaire-derived, examination-derived, and laboratory variables. Demographic variables include age, sex, marital status, and education. Examination and derived clinical variables include BMI, blood pressure, hypertension, smoking status, second-hand smoke exposure, and smoking duration.

Table A.10: Summary statistics (COPD).

Feature	Level	Type	Overall
BMXBMI		continuous	29.17 ± 6.98; missing 2127 (5.4%)
COPD	0	categorical	36675 (92.3%)
	1		3067 (7.7%)
DBQ229	1	categorical	15235 (38.3%)
	3		14841 (37.3%)
	2		9635 (24.2%)
	9		29 (0.1%)
	7		2 (0.0%)
DBQ700	3	categorical	16252 (40.9%)
	4		9381 (23.6%)
	2		8261 (20.8%)
	1		3508 (8.8%)
	5		2315 (5.8%)
	9		23 (0.1%)
	7		2 (0.0%)
DMDEDUC2	4	categorical	11561 (29.1%)
	3		9101 (22.9%)
	5		8934 (22.5%)
	2		5707 (14.4%)
	1		4380 (11.0%)
	9		41 (0.1%)

Continued on next page

Table A.10: Summary statistics (COPD).

Feature	Level	Type	Overall
	7		18 (0.0%)
DMDMARTL	1	categorical	20256 (51.0%)
	5		7234 (18.2%)
	3		4292 (10.8%)
	2		3369 (8.5%)
	6		3218 (8.1%)
	4		1339 (3.4%)
	77		30 (0.1%)
	99		4 (0.0%)
LBDBANO		continuous	0.05 ± 0.06; missing 3556 (8.9%)
LBDEONO		continuous	0.20 ± 0.18; missing 3556 (8.9%)
LBDMONO		continuous	0.56 ± 0.21; missing 3556 (8.9%)
LBDSGLSI		continuous	5.73 ± 2.21; missing 4004 (10.1%)
LBXHCT		continuous	41.41 ± 4.33; missing 3485 (8.8%)
LBXHGB		continuous	14.03 ± 1.56; missing 3485 (8.8%)
LBXRBCSI		continuous	4.66 ± 0.51; missing 3485 (8.8%)
LBXRDW		continuous	13.33 ± 1.38; missing 3485 (8.8%)
LBXSAL		continuous	4.21 ± 0.36; missing 4002 (10.1%)
LBXSBU		continuous	13.70 ± 6.12; missing 4009 (10.1%)
LBXSC3SI		continuous	25.09 ± 2.33; missing 4083 (10.3%)
LBXSCA		continuous	9.39 ± 0.37; missing 4037 (10.2%)
LBXSCLSI		continuous	103.44 ± 3.08; missing 4005 (10.1%)
LBXSGB		continuous	2.94 ± 0.46; missing 4057 (10.2%)
LBXSKSI		continuous	3.99 ± 0.35; missing 4011 (10.1%)
LBXSLDSI		continuous	134.12 ± 33.30; missing 4202 (10.6%)
LBXSNASI		continuous	139.31 ± 2.44; missing 4005 (10.1%)
LBXSOSSI		continuous	278.65 ± 5.46; missing 4011 (10.1%)
LBXSPH		continuous	3.73 ± 0.57; missing 4011 (10.1%)
LBXSTB		continuous	0.66 ± 0.32; missing 4027 (10.1%)
LBXSTP		continuous	7.15 ± 0.48; missing 4056 (10.2%)
LBXSTR		continuous	154.46 ± 130.48; missing 4027 (10.1%)
LBXSUA		continuous	5.44 ± 1.46; missing 4014 (10.1%)
LBXTC		continuous	193.12 ± 42.06; missing 3888 (9.8%)
LBXWBCSI		continuous	7.28 ± 3.23; missing 3486 (8.8%)
MCQ010	2	categorical	34071 (85.7%)
	1		5637 (14.2%)
	9		33 (0.1%)
	7		1 (0.0%)
MCQ160B	2	categorical	38280 (96.3%)
	1		1355 (3.4%)
	9		107 (0.3%)
MCQ160C	2	categorical	37940 (95.5%)
	1		1644 (4.1%)
	9		157 (0.4%)
	7		1 (0.0%)
NLR		continuous	2.19 ± 1.22; missing 3556 (8.9%)
RBC_FOLATE_NMOL		continuous	1113.39 ± 556.52; missing 5595 (14.1%)
RIAGENDR	2	categorical	20498 (51.6%)
	1		19244 (48.4%)
RIDAGEYR		continuous	49.71 ± 18.01

Continued on next page

Table A.10: Summary statistics (COPD).

Feature	Level	Type	Overall
SERUM_IRON		continuous	84.36 ± 35.58; missing 4024 (10.1%)
SHS	no	categorical	18725 (47.1%)
	Missing		12701 (32.0%)
	yes		8316 (20.9%)
SMOKING_STATUS	never	categorical	22077 (55.6%)
	past		9504 (23.9%)
	Missing		8161 (20.5%)
SMOKING_YEARS		continuous	11.36 ± 43.95; missing 2066 (5.2%)

A.4.6 Task-specific notes

Several derived variables are constructed during cohort generation. Mean systolic and diastolic blood pressure are computed from the first two available blood pressure measurements. Hypertension is derived from these mean blood pressure values. Smoking status, second-hand smoke exposure, smoking duration, neutrophil-to-lymphocyte ratio, red blood cell folate, and serum iron are also derived from NHANES questionnaire or laboratory variables.

The source paper did not specify how mean systolic and diastolic blood pressure were calculated. We therefore computed them from the first two available blood pressure measurements, matching the convention used in the depression prediction task.

Second-hand smoke exposure was derived from serum cotinine using the range 0.05–10.00 ng/mL, following the CDC/NCHS definition used for adult second-hand smoke exposure. The NHANES laboratory documentation describes serum cotinine as an indicator of tobacco-smoke exposure, while the CDC/NCHS Data Brief provides the specific range used to distinguish second-hand smoke exposure from no detected exposure and likely active smoking [55], [56]. This was done because serum cotinine is available across all the included NHANES cycles.

The original study's cycles were extended from 2009-2018

A.5 Depression

A.5.1 Task summary

This task predicts depression in adult NHANES participants. The cohort consists of participants aged 18 years or older from the 2007 through 2016 NHANES survey cycles.

A.5.2 Prediction target

The prediction target is a binary depression label. Depression is derived from the Patient Health Questionnaire-9 (PHQ-9), computed as the sum of the nine depres-

sion questionnaire items. A participant is labeled as positive if their PHQ-9 score is at least 10, and negative otherwise. Participants with missing or invalid PHQ-9 scores are excluded.

A.5.3 Input table and unit of observation

The input data is derived from NHANES survey, examination, questionnaire, and laboratory files. Data from multiple survey cycles are combined into a single table. Each row in the final task table corresponds to one NHANES participant, identified by survey cycle and participant identifier.

A.5.4 Cohort definition

Participants are included if they satisfy all of the following criteria:

- participation in one of the included NHANES survey cycles from 2007 to 2016.
- age of at least 18 years.
- sufficient questionnaire information to compute the PHQ-9 score.

Participants with missing or invalid PHQ-9 scores after applying the above definition are excluded.

A.5.5 Features

The feature table contains demographic, socioeconomic, examination-derived, lifestyle, and clinical variables. Demographic and socioeconomic variables include age, sex, ethnicity, education, marital status, and poverty income ratio.

Examination-derived and clinical variables include BMI, mean systolic and diastolic blood pressure, estimated glomerular filtration rate (eGFR), hypertension, diabetes, and dyslipidemia. Lifestyle variables include smoking status, drinking status, and physical activity.

Table A.11: Summary statistics (Depression).

Feature	Level	Type	Overall
BMXBMI		continuous	29.04 ± 6.98; missing 281 (1.1%)
BPXDIMEAN		continuous	69.57 ± 12.90; missing 608 (2.3%)
BPXSYMEAN		continuous	123.71 ± 18.45; missing 608 (2.3%)
DEPRESSION	0	categorical	23980 (90.9%)
	1		2412 (9.1%)
DIABETES	no	categorical	22046 (83.5%)
	yes		4346 (16.5%)
DMDEDUC2	4	categorical	7325 (27.8%)
	5		5765 (21.8%)
	3		5696 (21.6%)
	2		3653 (13.8%)

Continued on next page

Table A.11: Summary statistics (Depression).

Feature	Level	Type	Overall
	1		2562 (9.7%)
	Missing		1373 (5.2%)
	9		12 (0.0%)
	7		6 (0.0%)
DMDMARTL	1	categorical	12785 (48.4%)
	5		4628 (17.5%)
	3		2774 (10.5%)
	6		1996 (7.6%)
	2		1979 (7.5%)
	Missing		1373 (5.2%)
	4		844 (3.2%)
	77		12 (0.0%)
	99		1 (0.0%)
DRINKING_STATUS	current	categorical	15393 (58.3%)
	never		7554 (28.6%)
	past		2872 (10.9%)
	Missing		573 (2.2%)
DYSLIPIDEMIA	yes	categorical	17350 (65.7%)
	no		8023 (30.4%)
	Missing		1019 (3.9%)
HYPERTENSION	no	categorical	16574 (62.8%)
	yes		9418 (35.7%)
	Missing		400 (1.5%)
INDFMPIR		continuous	2.45 ± 1.63; missing 2244 (8.5%)
PHYSICAL_ACTIVITY	moderate	categorical	16073 (60.9%)
	mild		8830 (33.5%)
	vigorous		1481 (5.6%)
	Missing		8 (0.0%)
RIAGENDR	2	categorical	13370 (50.7%)
	1		13022 (49.3%)
RIDAGEYR		continuous	47.99 ± 18.56
RIDRETH1	3	categorical	11076 (42.0%)
	4		5611 (21.3%)
	1		4162 (15.8%)
	2		2824 (10.7%)
	5		2719 (10.3%)
SMOKING_STATUS	never	categorical	14228 (53.9%)
	Missing		6092 (23.1%)
	past		6072 (23.0%)
eGFR		continuous	95.85 ± 23.18; missing 1376 (5.2%)

A.5.6 Task-specific notes

Several derived variables are constructed during cohort generation. Mean systolic and diastolic blood pressure are computed from the first two available blood pressure measurements. Smoking status, drinking status, hypertension, diabetes, and dyslipidemia are derived from NHANES questionnaire, examination, or laboratory variables.

The PHQ-9 score is computed only when all nine questionnaire items have valid

response values. Invalid, refused, unknown, or missing item responses make the PHQ-9 score invalid, and those participants are excluded from the task.

The source paper calculated eGFR using the MDRD equation. In this implementation, eGFR was instead calculated using the 2021 race-free CKD-EPI[57] creatinine equation to align the derived kidney-function feature with the other NHANES tasks in the benchmark. This choice was made for cross-task consistency and should be interpreted as an implementation adaptation rather than an exact reproduction of the source-paper feature definition.

The source paper did not specify how physical activity was operationalized. We therefore constructed a best-effort categorical variable from the available NHANES physical-activity questionnaire items, distinguishing vigorous, moderate, and mild activity. Refused, unknown, or missing responses were treated as missing.

The source paper did not fully specify how physical activity, smoking status, or drinking status were operationalized. We therefore constructed best-effort categorical variables from the available NHANES questionnaire items. Physical activity was derived from work-related and recreational activity questions. Smoking status was derived from lifetime smoking history and time since quitting. Drinking status was derived from lifetime alcohol use and reported drinking frequency over the past 12 months. Refused, unknown, or otherwise invalid responses were treated as missing.

The original study's included cycle was expanded from only 2013-2014 to 2007-2016

A.6 Diabetes

A.6.1 Task summary

This task predicts diabetes status in adult NHANES participants. Unlike several other NHANES tasks, it is not based on a single source-paper cohort definition. Instead, the task is defined directly from standard glycemic thresholds using NHANES laboratory measurements. The cohort consists of participants aged 18 years or older from the 2005 through 2023 NHANES survey cycles.

A.6.2 Prediction target

The prediction target is a three-class diabetes status label derived from glycohemoglobin and serum glucose measurements. Participants are labeled as diabetic if glycohemoglobin is at least 6.5% or serum glucose is at least 126 mg/dL. Participants are labeled as prediabetic if glycohemoglobin is at least 5.7% or serum glucose is at least 100 mg/dL, but neither diabetic threshold is met. Remaining participants are labeled as non-diabetic.

The diabetes thresholds follow standard diagnostic cut points for HbA1c and fasting glucose, while the intermediate class uses the ADA prediabetes thresholds [44]. Participants with missing glycohemoglobin or serum glucose measurements are excluded. The glycohemoglobin and serum glucose variables used to construct the

label are removed from the feature table to avoid target leakage.

A.6.3 Input table and unit of observation

The input data is derived from NHANES survey, examination, questionnaire, and laboratory files. Data from multiple survey cycles are combined into a single table. Each row in the final task table corresponds to one NHANES participant, identified by survey cycle and participant identifier.

A.6.4 Cohort definition

Participants are included if they satisfy all of the following criteria:

- participation in one of the included NHANES survey cycles from 2005 to 2023.
- age of at least 18 years.
- available glycohemoglobin measurement.
- available serum glucose measurement.

Participants with missing values for either measurement used to derive the diabetes status label are excluded.

A.6.5 Features

The feature table contains demographic, socioeconomic, questionnaire-derived, examination-derived, and laboratory variables. Demographic and socioeconomic variables include age, sex, ethnicity, education, and poverty income ratio.

Examination-derived variables include BMI, waist circumference, systolic blood pressure, and diastolic blood pressure. Laboratory variables include triglycerides, HDL cholesterol, and total cholesterol. Questionnaire-derived variables include smoking-related variables.

Table A.12: Summary statistics (Diabetes).

Feature	Level	Type	Overall
BMXBMI		continuous	29.18 ± 7.08; missing 636 (1.4%)
BMXWAIST		continuous	99.05 ± 16.65; missing 2175 (4.7%)
BPXDI1		continuous	71.17 ± 12.92; missing 3580 (7.7%)
BPXSY1		continuous	123.68 ± 18.80; missing 3580 (7.7%)
DMDEDUC2	4	categorical	13118 (28.2%)
	5		10847 (23.3%)
	3		9990 (21.5%)
	2		5791 (12.4%)
	1		4318 (9.3%)
	Missing		2433 (5.2%)

Continued on next page

Table A.12: Summary statistics (Diabetes).

Feature	Level	Type	Overall
	9		34 (0.1%)
	7		11 (0.0%)
INDFMPIR		continuous	2.54 ± 1.64; missing 4551 (9.8%)
LBDHDD		continuous	53.30 ± 15.94; missing 10 (0.0%)
LBXSCH		continuous	190.87 ± 42.46; missing 42 (0.1%)
LBXSTR		continuous	147.56 ± 122.91; missing 27 (0.1%)
RIAGENDR	2	categorical	24078 (51.7%)
	1		22464 (48.3%)
RIDAGEYR		continuous	48.54 ± 18.67
RIDRETH1	3	categorical	20174 (43.3%)
	4		9538 (20.5%)
	1		6910 (14.8%)
	5		5329 (11.4%)
	2		4591 (9.9%)
SMD415A	Missing	categorical	19061 (41.0%)
	0		17254 (37.1%)
	1		6381 (13.7%)
	2		3206 (6.9%)
	3		614 (1.3%)
	777		16 (0.0%)
	999		10 (0.0%)
SMD680	2	categorical	32118 (69.0%)
	1		10431 (22.4%)
	Missing		3987 (8.6%)
	7		4 (0.0%)
	9		2 (0.0%)
SMQ020	2	categorical	25718 (55.3%)
	1		19478 (41.9%)
	Missing		1311 (2.8%)
	9		26 (0.1%)
	7		9 (0.0%)
outcome	0	categorical	23835 (51.2%)
	1		15950 (34.3%)
	2		6757 (14.5%)

A.6.6 Task-specific notes

This task is a benchmark-defined NHANES task rather than a reproduction of a source-paper task. The outcome is derived from laboratory thresholds for glycohemoglobin and serum glucose. The variables used to construct the outcome, `LBXGH` and `LBXSGL`, are excluded from the predictors to avoid direct leakage.

The three outcome classes are clinically ordered as non-diabetic, prediabetic, and

diabetic. However, the benchmark treats the task as standard multiclass classification.

A.7 HbA1c

A.7.1 Task summary

This task predicts glycohemoglobin (HbA1c) in adult NHANES participants[58]. The cohort consists of participants aged 18 years or older from the 2005 through 2023 NHANES survey cycles.

A.7.2 Prediction target

The prediction target is the continuous HbA1c value. The outcome is derived from the `LBXGH` laboratory variable. Participants with missing HbA1c measurements are excluded.

A.7.3 Input table and unit of observation

The input data is derived from NHANES survey, examination, and laboratory files. Data from multiple survey cycles are combined into a single table. Each row in the final task table corresponds to one NHANES participant, identified by survey cycle and participant identifier.

A.7.4 Cohort definition

Participants are included if they satisfy all of the following criteria:

- participation in one of the included NHANES survey cycles from 2005 to 2023.
- age of at least 18 years.
- available HbA1c measurement.

Participants with missing HbA1c values after applying the above definition are excluded.

A.7.5 Features

The feature table contains demographic, examination-derived, and laboratory variables. Demographic variables include age and education. Examination-derived variables include waist circumference, systolic blood pressure, BMI, body weight, arm circumference, and upper leg length.

Laboratory variables include serum glucose, osmolality, triglycerides, chloride, albumin, blood urea nitrogen, globulin, urinary albumin-to-creatinine ratio, sodium, HDL cholesterol, and urinary albumin.

Table A.13: Summary statistics (HbA1c).

Feature	Level	Type	Overall
BMXARMC		continuous	33.15 ± 5.23; missing 1752 (3.7%)
BMXBMI		continuous	29.18 ± 7.10; missing 668 (1.4%)
BMXLEG		continuous	38.71 ± 3.89; missing 2473 (5.2%)
BMXWAIST		continuous	99.06 ± 16.68; missing 2267 (4.8%)
BMXWT		continuous	81.68 ± 21.91; missing 580 (1.2%)
BPXSY1		continuous	123.75 ± 18.89; missing 3682 (7.7%)
BPXSY2		continuous	123.09 ± 18.52; missing 16944 (35.6%)
BPXSY3		continuous	122.35 ± 18.17; missing 17087 (35.9%)
DMDEDUC2	4	categorical	13392 (28.2%)
	5		11065 (23.3%)
	3		10242 (21.5%)
	2		5919 (12.4%)
	1		4409 (9.3%)
	Missing		2486 (5.2%)
	9		35 (0.1%)
	7		11 (0.0%)
LBDHDDSI		continuous	1.38 ± 0.41; missing 843 (1.8%)
LBDSALSI		continuous	41.91 ± 3.67; missing 1008 (2.1%)
LBDSBUSI		continuous	4.93 ± 2.16; missing 1051 (2.2%)
LBDSGBSI		continuous	29.63 ± 4.57; missing 1065 (2.2%)
LBDSGLSI		continuous	5.68 ± 2.15; missing 1017 (2.1%)
LBDSTRSI		continuous	1.67 ± 1.39; missing 1037 (2.2%)
LBXGH		continuous	5.74 ± 1.08
LBXSCLSI		continuous	103.10 ± 3.12; missing 1045 (2.2%)
LBXSNASI		continuous	139.43 ± 2.45; missing 1011 (2.1%)
LBXSOSSI		continuous	278.88 ± 5.46; missing 1057 (2.2%)
RIDAGEYR		continuous	48.62 ± 18.69
URDACT		continuous	43.62 ± 327.53; missing 11201 (23.6%)
URXUMA		continuous	45.53 ± 326.36; missing 761 (1.6%)

A.7.6 Task-specific notes

This task is formulated as a regression problem. The HbA1c variable `LBXGH` is used only as the prediction target and is excluded from the feature table to avoid direct leakage.

This task is a partial reconstruction of the source study rather than an exact reproduction. The source study selected features from a broad NHANES variable pool using a data-driven informativeness criterion, but the published description did not provide a fully unambiguous mapping for all selected variables. We therefore used only the identifiable de-duplicated variables from the reported feature table. Variables that appeared to duplicate another feature through unit conversion, lacked a clear NHANES code, or were not consistently available across the included cycles

were excluded.

The original study's included cycles were expanded from 2007-2020 to 2005-2023.

A.8 Hypertension

A.8.1 Task summary

This task predicts hypertension status in adult NHANES participants[32]. The cohort consists of participants aged 20 years or older from the 2001 through 2023 NHANES survey cycles.

A.8.2 Prediction target

The prediction target is a binary hypertension label. A participant is labeled as positive if their first systolic blood pressure reading is at least 130 mmHg, and negative otherwise. Participants without the required systolic blood pressure measurement are excluded.

A.8.3 Input table and unit of observation

The input data is derived from NHANES survey, examination, questionnaire, and laboratory files. Data from multiple survey cycles are combined into a single table. Each row in the final task table corresponds to one NHANES participant, identified by survey cycle and participant identifier.

A.8.4 Cohort definition

Participants are included if they satisfy all of the following criteria:

- participation in one of the included NHANES survey cycles from 2001 to 2023.
- age of at least 20 years.
- available first systolic blood pressure reading.

Participants with missing hypertension labels after applying the above definition are excluded.

A.8.5 Features

The feature table contains demographic, examination-derived, questionnaire-derived, and derived clinical variables. Demographic variables include age, sex, and ethnicity. Examination-derived variables include BMI. Questionnaire-derived variables include smoking status and diabetes status. Chronic kidney disease is included as a derived clinical variable.

Table A.14: Summary statistics (Hypertension).

Feature	Level	Type	Overall
BMXBMI		continuous	29.10 ± 6.84; missing 740 (1.4%)
CKD	0	categorical	48836 (94.0%)
	Missing		1687 (3.2%)
	1		1406 (2.7%)
DIQ010	2	categorical	44144 (85.0%)
	1		6562 (12.6%)
	3		1193 (2.3%)
	9		30 (0.1%)
RIAGENDR	2	categorical	26731 (51.5%)
	1		25198 (48.5%)
RIDAGEYR		continuous	50.05 ± 17.99
RIDRETH1	3	categorical	23636 (45.5%)
	4		10710 (20.6%)
	1		7742 (14.9%)
	5		5257 (10.1%)
	2		4584 (8.8%)
SMQ020	2	categorical	28559 (55.0%)
	1		23330 (44.9%)
	9		28 (0.1%)
	7		9 (0.0%)
	Missing		3 (0.0%)
outcome	0	categorical	34090 (65.6%)
	1		17839 (34.4%)

A.8.6 Task-specific notes

The hypertension label is derived from the first systolic blood pressure reading using a threshold of 130 mmHg.

Chronic kidney disease is derived differently depending on survey cycle. For earlier cycles where the kidney-function questionnaire item is available, CKD is derived from self-reported weak or failing kidneys. For later cycles, CKD is derived from estimated glomerular filtration rate and urinary albumin–creatinine ratio. The eGFR variable is calculated using the 2021 race-free CKD-EPI creatinine equation to align kidney-function feature construction with the other NHANES tasks in the benchmark.

The included survey cycles were expanded beyond the original 2007-2016 to 2001-2023 to increase the number of available participants and to align this task with the broader NHANES benchmark setup.

A.9 MASLD

A.9.1 Task summary

This task predicts metabolic dysfunction-associated steatotic liver disease (MASLD) in adult NHANES participants[34]. The cohort consists of participants aged 20 years or older from the 2017 to 2023 NHANES survey cycles.

A.9.2 Prediction target

The prediction target is a binary MASLD label. MASLD is derived from transient elastography controlled attenuation parameter (CAP). A participant is labeled as positive if their median CAP value is at least 302 dB/m, and negative otherwise. Participants without a valid CAP measurement are excluded.

A.9.3 Input table and unit of observation

The input data is derived from NHANES survey, examination, questionnaire, laboratory, and liver ultrasound transient elastography files. Data from multiple survey cycles are combined into a single table. Each row in the final task table corresponds to one NHANES participant, identified by survey cycle and participant identifier.

A.9.4 Cohort definition

Participants are included if they satisfy all of the following criteria:

- participation in one of the included NHANES survey cycles from 2017 to 2023.
- age of at least 20 years.
- available median CAP measurement.
- no positive hepatitis B or hepatitis C marker.
- no excluded alcohol-consumption category.
- complete data for all retained predictors and the MASLD label.

Participants with missing values in the final selected feature set are excluded.

A.9.5 Features

The feature table contains demographic, examination-derived, and laboratory variables. Demographic variables include age, sex, ethnicity, and education.

Examination-derived variables include waist circumference, BMI, systolic blood pressure, and diastolic blood pressure. Laboratory variables include triglycerides, ALT, HDL cholesterol, LDL cholesterol, total cholesterol, alkaline phosphatase, AST, blood urea nitrogen, creatine phosphokinase, globulin, lactate dehydrogenase, total bilirubin, uric acid, C-reactive protein, serum glucose, and gamma-glutamyl transferase.

Table A.15: Summary statistics (MASLD).

Feature	Level	Type	Overall
BMXBMI		continuous	30.31 ± 7.23
BMXWAIST		continuous	102.23 ± 17.13
DBP		continuous	74.68 ± 11.39
DMDEDUC2	4	categorical	1080 (31.6%)
	5		908 (26.5%)
	3		843 (24.6%)
	2		341 (10.0%)
	1		248 (7.3%)
LBDHDDSI		continuous	1.36 ± 0.37
LBDLDLSI		continuous	2.83 ± 0.94
LBDSGBSI		continuous	30.98 ± 4.19
LBDSGLSI		continuous	5.96 ± 2.07
LBDSTBSI		continuous	8.52 ± 4.95
LBDSUASI		continuous	316.39 ± 84.14
LBXHSCRP		continuous	4.21 ± 8.14
LBXSAPSI		continuous	81.81 ± 26.89
LBXSASSI		continuous	21.45 ± 13.99
LBXSATSI		continuous	21.65 ± 18.40
LBXSBU		continuous	15.04 ± 5.78
LBXSCH		continuous	185.50 ± 41.33
LBXSCK		continuous	148.43 ± 385.38
LBXSGTSI		continuous	29.23 ± 51.43
LBXSLDSI		continuous	159.93 ± 34.64
LBXSTR		continuous	122.03 ± 64.34
MASLD	0	categorical	2391 (69.9%)
	1		1029 (30.1%)
RIAGENDR	2	categorical	1852 (54.2%)
	1		1568 (45.8%)
RIDAGEYR		continuous	53.58 ± 16.81
RIDRETH1	3	categorical	1579 (46.2%)
	4		613 (17.9%)
	5		453 (13.2%)
	1		404 (11.8%)
	2		371 (10.8%)
SBP		continuous	123.60 ± 18.98

A.9.6 Task-specific notes

Systolic and diastolic blood pressure are derived from the first available blood pressure reading. Since availability of codes differs across the included cycles, the first reading is used consistently.

Unlike several other NHANES tasks, this task applies complete-case filtering after

feature construction. Participants with missing values in any retained feature are excluded from the final task table.

The included survey cycles were expanded beyond the original 2017-2020 to 2017-2023 to increase the number of available participants, and to enable a temporal train/test split.

A.10 NAFLD

A.10.1 Task summary

This task predicts non-alcoholic fatty liver disease (NAFLD) in adolescent NHANES participants. The cohort consists of participants aged 12 to 19 years from the 2009 through 2023 NHANES survey cycles.

A.10.2 Prediction target

The prediction target is a binary NAFLD label. The label is derived from alanine aminotransferase (ALT). Male participants are labeled as positive if ALT is greater than 26 U/L, and female participants are labeled as positive if ALT is greater than 22 U/L. Participants below or equal to the sex-specific threshold are labeled as negative.

A.10.3 Input table and unit of observation

The input data is derived from NHANES survey, examination, and laboratory files. Data from multiple survey cycles are combined into a single table. Each row in the final task table corresponds to one NHANES participant, identified by survey cycle and participant identifier.

A.10.4 Cohort definition

Participants are included if they satisfy all of the following criteria:

- participation in one of the included NHANES survey cycles from 2009 to 2023.
- age between 12 and 19 years, inclusive.
- available triglyceride, fasting glucose, and ALT measurements.
- available waist circumference, height, weight, and BMI measurements.

Participants missing any of the required measurements above are excluded.

A.10.5 Features

The feature table contains demographic, examination-derived, and laboratory variables. Demographic variables include age, sex, and ethnicity.

Examination-derived variables include waist circumference, height, weight, and BMI. Laboratory variables include triglycerides, platelet count, fasting glucose, LDL cholesterol, white blood cell count, total cholesterol, red blood cell count, HDL cholesterol, hemoglobin, and glycohemoglobin.

Table A.16: Summary statistics (NAFLD).

Feature	Level	Type	Overall
BMXBMI		continuous	24.60 ± 6.63
BMXHT		continuous	165.55 ± 10.03
BMXWAIST		continuous	83.09 ± 16.12
BMXWT		continuous	67.93 ± 20.87
LBDHDD		continuous	52.39 ± 11.77; missing 1 (0.0%)
LBDLDL		continuous	86.81 ± 25.37; missing 39 (1.7%)
LBXGH		continuous	5.26 ± 0.45; missing 2 (0.1%)
LBXGLU		continuous	95.64 ± 14.57
LBXHGB		continuous	13.98 ± 1.37; missing 3 (0.1%)
LBXPLTSI		continuous	256.12 ± 57.16; missing 3 (0.1%)
LBXRBCSI		continuous	4.84 ± 0.45; missing 3 (0.1%)
LBXSCH		continuous	156.12 ± 29.27
LBXSTR		continuous	80.75 ± 48.32
LBXWBCSI		continuous	6.35 ± 1.86; missing 3 (0.1%)
NAFLD	0	categorical	1948 (85.8%)
	1		322 (14.2%)
RIAGENDR	1	categorical	1169 (51.5%)
	2		1101 (48.5%)
RIDAGEYR		continuous	15.36 ± 2.21
RIDRETH1	3	categorical	733 (32.3%)
	4		511 (22.5%)
	1		484 (21.3%)
	5		322 (14.2%)
	2		220 (9.7%)

A.10.6 Task-specific notes

The original study provided a processed dataset, but we were unable to reproduce the reported cohort and feature table from the available files and documentation. We therefore reconstructed the task directly from NHANES source files using the identifiable variables and outcome definition described above.

The source task included hepatitis-based exclusion criteria, but the corresponding NHANES variables could not be identified unambiguously from the available task description. These exclusions were therefore not applied in this implementation. This choice should be interpreted as an implementation adaptation rather than an exact reproduction of the source task.

Unlike the MASLD task, this task is restricted to adolescent participants and uses

ALT-based labeling rather than transient elastography CAP-based labeling.

The original study states that the included ages were 11-20 years old. Upon examining the provided data, we concluded that this was a typing error, and should be 12-19.

The included survey cycles were expanded beyond the original 2011-2020 to 2009-2023 to increase the number of available participants. [35]