Case-Control Matching Erodes Feature Discriminability for AI-driven Sepsis Prediction in ICUs: A Retrospective Cohort Study

Sophia F. Ehlers[1,2], Fanny Tranchellini[2], Youssef Farag[2], Tim Hahn[1], Catherine Jutzeler[†2,3], and Lakmal Meegahapola[†2,3]

[1]Department of Medicine, University of Münster, Münster, Germany
[2]Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland
[3]Swiss Institute for Bioinformatics, Lausanne, Switzerland

ABSTRACT

Background: Sepsis remains a leading cause of intensive care unit (ICU) mortality worldwide, and early detection is essential for improving survival through timely interventions. While machine learning holds promise for early sepsis detection by leveraging the multimodal time series data prevalent in ICUs, the field faces significant methodological challenges—such as class imbalance and temporal misalignment. These issues have spurred growing interest in case-control matching strategies. However, there is a lack of systematic evaluation of how commonly used case-control matching methods affect model performance and introduce potential biases, highlighting a critical gap in current knowledge.

Methods: We investigated how matched training data affect both the performance of various machine learning architectures and the predictive importance of individual clinical features. Three harmonized large-scale ICU cohorts were used: the high time-resolution ICU dataset (HiRID) from Bern University Hospital, Switzerland (29,698 stays, 2008–2019), with a sepsis prevalence of 6.3%; the Medical Information Mart for Intensive Care (MIMIC-IV) database from Beth Israel Deaconess Medical Center, USA (63,425 stays, 2008–2019), with a sepsis prevalence of 5.2%; and the eICU Collaborative Research Database from 208 US hospitals (123,413 stays, 2014–2015), with a sepsis prevalence of 4.6%. Each dataset included hourly observations. both with and without incorporating demographic variables, and compared the results to unmatched and undersampled cohorts with equivalent ratios. To evaluate how matching strategies influenced feature relevance, we applied a Linear Mixed Effects Model (LMEM) to assess changes in the predictive significance of individual features. Finally, we trained machine learning models—including random forests and gradient-boosted trees—and evaluated their performance on the original test sets using the AUROC and normalized AUPRC metrics.

Findings: The LMEM results demonstrated that case-control matching led to a substantial reduction in the number of features identified as significant, based on multiple testing–corrected p-values, across all three cohorts. Specifically, the number of significant features declined from 35–43 in the original datasets to 24–29 as the matching ratio increased from 1:10 to 1:2. In the machine learning experiments, models trained on undersampled data achieved strong performance, with AUROC values exceeding 0.90 and normalized AUPRC scores above 41. In contrast, models trained on the original (imbalanced) datasets performed robustly (AUROC up to 0.82, normalized AUPRC up to 4.2), whereas those trained on matched datasets showed significant performance degradation, with AUROC values below 0.50 and normalized AUPRC scores falling below baseline. These patterns were consistent across all three cohorts and a wide range of evaluation setups.

Interpretation: This multi-cohort analysis highlights a methodological paradox: while case-control matching can mitigate class imbalance and help disentangle sepsis-specific patterns from general ICU trajectories, overly strict matching criteria may significantly impair predictive performance. These findings underscore the need for more nuanced matching strategies that balance bias reduction with the preservation of critical clinical signals—ultimately enhancing the reliability and real-world applicability of sepsis prediction models.

Funding: This project was supported by grant #902 of the Strategic Focus Area “Personalized Health and Related Technologies (PHRT)” of the ETH Domain and Young Investigator Grant of the Novartis Foundation for Medical-Biological Research.

Research in Context

Evidence before this study: Case-control matching has long been used in epidemiological research to mitigate confounders and address class imbalance. In recent years, AI-driven sepsis prediction studies have adopted similar strategies to improve data set balance and reduce temporal bias. However, the direct impact of case-control matching on machine learning (ML) performance in sepsis prediction has not been systematically evaluated. Although prior work demonstrated the promise of early detection of sepsis using observational data without explicit matching strategies[1,2], studies that implemented stringent matching occasionally reported deteriorations in model performance[3,4]. A recent review also advocated the use of case-control temporal alignment to avoid temporal biases[2]. This discrepancy highlights a critical knowledge gap about how various matching protocols influence ML outcomes in sepsis, an inherently heterogeneous condition shaped by both clinical and temporal complexities.

Added value of this study: We conducted a comprehensive multi-cohort investigation of absolute-onset case-control matching (with and without demographic criteria) for sepsis prediction models, using three large, harmonized ICU cohorts (HiRID, MIMIC-IV, and eICU). By comparing matched, aligned, and undersampled training cohorts across multiple machine learning architectures and temporal window configurations, our experiments show how matching affects both overall model performance and the discriminability of individual features. We demonstrated that while case-control matching effectively reduces class imbalance, it can also diminish key sepsis-related signals and substantially reduce predictive accuracy below baseline. These insights shed light on the nuanced interaction between epidemiological design strategies and ML-based sepsis prediction in ICU settings.

Implications of all the available evidence: Although case-control matching is often viewed as a reliableethod to minimize bias and improve dadata setbalance, our findings indicate that it may inadvertently obscure clinically significant signals critical for accurate prediction of sepsis.is. In real-world clinical environments where timely and precise diagnosis is essential, researchers and clinicians should carefully consider the trade-offs imposed by strict matching protocols, especially when aligning data from large multicenter cohorts. To facilitate transparent model validation, we recommend using absolute-onset matching as a comparative benchmark to prevent artificially inflated performance metrics. Alternatively, temporal dependencies related to admission to the ICU or matched onset should be factored into the model evaluation to avoid trivial classification driven by inadequate handling of within-stay dynamics. Lastly, relying solely on AUROC and AUPRC as performance metrics risks overlooking other clinically relevant outcomes, emphasizing the need for broader evaluation frameworks that align more closely with real-world clinical decision making. This study encourages the research community to refine and reassess conventional matching methods, striving for clinically informed strategies that reduce confounding while retaining essential sepsis-specific signals for robust and generalizable predictive models.

[†]Shared last authorship.

1 Introduction

Sepsis, a life-threatening condition characterized by organ dysfunction resulting from a dysregulated host response to infection, remains a major contributor to mortality in intensive care units (ICUs) worldwide[2]. Despite advances in antimicrobial therapies and critical care practices, in-hospital mortality rates for sepsis continue to exceed 20%[5]. Early detection and timely intervention are critical, as prompt administration of appropriate antibiotics is among the most effective strategies for improving surviva[6]. Notably, each hour of delay in antibiotic treatment is associated with a significant increase in mortality risk. This time-critical nature of sepsis management has spurred significant interest in predictive tools capable of alerting clinicians to early signs of deterioration—often before overt clinical symptoms emerge[7,8]. The application of artificial intelligence (AI), in particular machine learning (ML), has emerged as a promising avenue for advancing sepsis prediction in critical care settings. ML models, including deep neural networks and robust ensemble tree methods, are capable of analyzing the high-dimensional, multimodal data found in electronic health records (EHRs) to identify subtle patterns that may precede clinical recognition of sepsis[9].Interestingly, while deep learning techniques have made considerable strides in ICU prediction tasks, recent studies suggest that ensemble methods such as boosting algorithms and random forests often outperform deep learning models in practical applications[10–13]. Moreover, these models have demonstrated the ability to predict sepsis onset as early as 12–24 hours in advance—and in some cases, even earlier—prior to clinical recognition by healthcare providers. [14,15]. Such early warning systems could enable timely clinical interventions, illustrating the substantial impact AI can have in enhancing decision-making and outcomes in acute care settings.

Despite recent advances, translating AI models for sepsis prediction into reliable and impactful bedside tools remains a complex challenge. A fundamental obstacle lies in the inherent heterogeneity of critical care environments, which differ substantially across hospitals and regions due to variations in patient populations, data collection frequencies, and clinical protocols[16]. Additionally, the definition of sepsis has evolved over time, with the current Sepsis-3 criteria defining it as an increase of over $\geq$2 points in the Sequential Organ Failure Assessment (SOFA) score in the presence of suspected infection[17]. However, the practical application of these criteria remains inconsistent, potentially introducing noise and bias in outcome labeling. One of the most significant challenges, even within a single cohort, is the pronounced class imbalance inherent in ICU data, where sepsis cases are vastly outnumbered by non-septic admissions. This imbalance is often compounded by differences in ICU length of stay and clinical trajectories between septic and non-septic patients, increasing the risk that models will learn spurious or non-causal associations[2]. For example, a model might erroneously associate longer ICU stays with sepsis, simply because septic patients are more likely to remain hospitalized for extended periods. Similarly, it may overemphasize age-related risk without capturing the underlying physiological mechanisms. These factors collectively complicate the development of robust, generalizable AI models for early sepsis detection.

One proposed methodological approach to address the challenges of class imbalance and temporal and demographic biases in sepsis prediction is case-control matching, an epidemiological technique that has been adapted for training ML models. In this approach, each identified sepsis case is paired with one or more non-sepsis control patients who exhibit similar baseline characteristics, such as age and pre-existing conditions, and who have a comparable duration of observation within the ICU[2]. By selectively sampling control patients to create a cohort that mirrors the sepsis cases in terms of key confounding variables and the alignment of the time period at risk, matching aims to generate a more balanced dataset. The theoretical advantage of this technique is that it should facilitate the ability of an algorithm to discern patterns specifically related to sepsis by mitigating obvious sources of bias, such as disparities in the length of hospital stays. For example, matching based on the time of sepsis onset after ICU admission, also known as "absolute onset" matching, ensures that a control patient was still present in the ICU at the time when the matched sepsis patient developed the condition, thereby preventing a trivial distinction based solely on the duration of their stay, and its' effect on the captured features[1,2,18]. The application of case-control matching has become increasingly common in recent sepsis studies seeking to address these inherent data complexities. Beyond temporal alignment, case-control matching offers opportunities to reduce confounding factors and balance distribution shifts of key covariates between cases and controls[19]. For instance, epidemiological research frequently employs case-control matching algorithms based on key patient characteristics, such as demographic features including sex and age[19,20]. This approach is potentially crucial because older patients generally face higher risks for numerous pathologies and clinical complications,

including sepsis. Failure to balance training data with respect to age might yield biased outcomes, with prediction models potentially identifying age rather than disease onset. On the contrary, a model should not predict sepsis just because of the age, but should also learn underlying pathophysiologies related to sepsis, beyond age[21]. However, it is important to recognize that matching can introduce selection bias, and the efficiency gains from matching are context-dependent and not universally guaranteed[21]. However, this approach has not consistently enhanced model robustness, as previous research has documented performance degradation in MGP-RNN models following case-control alignment[3,4]. This finding could imply that overly rigid matching protocols might eliminate valuable variability within the data, leading to the creation of an artificial patient cohort that, while statistically more homogeneous, is actually more challenging for an AI model to learn from effectively. In essence, the very process intended to increase precision in the selection of the study cohort could paradoxically reduce the discriminability of the relevant clinical features. These considerations have generated conflicting perspectives regarding the merits, necessity, and limitations of case-control matching and temporal alignment in sepsis prediction, despite the absence of structured, comprehensive examinations of this topic. Current literature lacks standardized reporting procedures for these methodological aspects, and significant knowledge gaps persist concerning the interaction between onset matching and various machine learning architectures.

This study aims to address this critical methodological challenge through a systematic analysis across multiple large-scale, diverse, and harmonized ICU cohorts. We hypothesize that while case-control matching can help control for confounding variables, its stringent application may inadvertently attenuate the predictive signal of clinically relevant features, ultimately compromising model performance. To investigate this, we conduct a comprehensive evaluation of absolute-onset case-control matching—applied with and without demographic feature alignment—and its impact on machine learning–based sepsis prediction. Our objectives are twofold. First, we assess how case-control matching influences the statistical significance and predictive contribution of individual clinical features, using rigorous feature-level analysis. Second, we evaluate how varying matching ratios affect the performance of machine learning models, specifically tree-based and gradient-boosted architectures. We compare models trained on matched cohorts to those trained on unmatched and undersampled datasets with equivalent case-to-control ratios, thereby isolating the effects of the matching strategy itself. Through this investigation, we aim to deepen our understanding of the complex interplay between epidemiological matching techniques and machine learning methods, and to provide guidance for the design of more robust and generalizable predictive models for early sepsis detection.

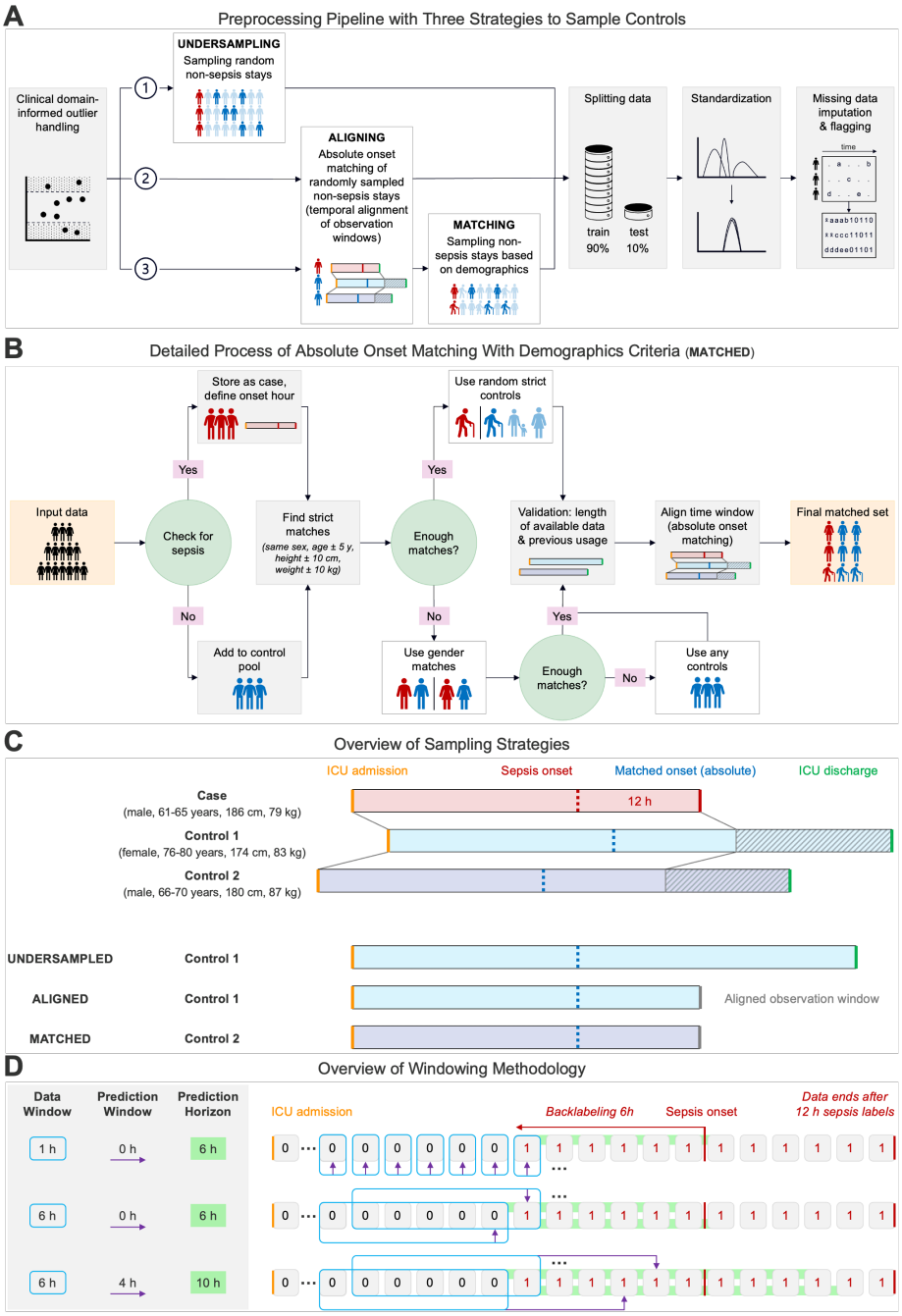


Figure 1. Case-control matching and windowing methodology. A) Preprocessing pipeline comprising outlier management, three case-control balancing strategies (1. undersampling, 2. alignment, and 3. matching – incorporating temporal alignment with demographics-based matching), followed by data splitting, standardization, and missing value imputation. **B)** Detailed process of absolute onset matching using demographics ("matched", A3). Sepsis cases are identified and time to onset determined. Controls are selected using a hierarchical matching approach: primary matching within predefined demographic parameter thresholds, secondary matching based on biological sex, and tertiary random assignment when other matching criteria could not be satisfied. Each control subject is utilized only once, with temporal data windows aligned between matched case-control pairs. **C)** Schematic illustration of absolute onset matching implementation with and without demographic variables, contrasted with undersampling methodology lacking temporal alignment. Mentioned demographic information are hypothetical, done for this illustration. *Absolute* refers to the absolute time since ICU admission. Controls must have monitoring duration equal to or exceeding their matched cases. All data collected beyond 12 hours post-matched absolute onset are excluded from analysis. **D)** Methodological Overview of temporal windowing with implementation of three data-prediction-window pairings: 1-hour/0-hours, 6-hours/0-hours, and 6-hours/4-hours. Positive sepsis labels are initiated six hours before officially defined sepsis onset ("backlabeling") and persist for 12 hours total, penalizing late false-negative predictions during model training. This approach yields a 6-hour prediction horizon when no prediction window (0 hours) is applied. Predictive assessments occur at hourly intervals.

Dataset		HiRID	MIMIC-IV	eICU
Originally published		2020 (Hyland, 2020)[12]	2020 (Johnson, 2020)[25]	2017 (Pollard, 2018)[27]
Version		v1.1.1	v2.0	v2.0
Years of Data Collection		2011-2016	2008-2019	2014-2015
Origin		Switzerland	USA	USA
Frequency (time-series)		2/5 minutes	1 hour	5 minutes
Number of patients		33,905	53,090	166,916
Number of ICU stays		32,338	75,652	182,774
Admission type	Medical	–	49,217 (65%)	134,632 (79%)
	Surgical	–	25,674 (34%)	51,909 (19%)
	Other	–	761 (1%)	4,702 (3%)
Age at admission (median [IQR])		65 [55, 75]	65 [53, 76]	65 [53, 76]
Gender	Male	20,796 (64%)	42,153 (56%)	98,834 (54%)
	Female	11,542 (36%)	33,499 (44%)	83,940 (46%)
Race	White	–	51,575 (76%)	140,958 (78%)
	Black	–	8,223 (12%)	19,887 (11%)
	Asian	–	2,225 (3%)	5,008 (3%)
	Other	–	5,514 (8%)	16,978 (9%)

Table 1. Details of ICU datasets HiRID, MIMIC-IV and eICU. These datasets were used for the YAIB harmonization workflow[22]. Each dataset is publicly available and can be accessed via Physionet after completing a credentialing procedure.

2 Methods

2.1 Harmonizing Datasets & Cohort Selection

This study used YAIB pipeline[22] and RiCU[23] to harmonize dataset and perform cohort selection. YAIB is a flexible multi-center framework designed for conducting reproducible and comparable machine learning experiments in intensive care settings. Our analysis focused specifically on three cohorts, that are natively supported by YAIB: the High Time Resolution ICU Dataset (HiRID)[12,24], the Medical Information Mart for Intensive Care IV (MIMIC-IV)[25,26], and the eICU Collaborative Research Database (eICU)[27,28]. More detailed provided in Table 1. The YAIB workflow initially focused on creating compatible datasets through unit harmonization, artifact removal, and variable mapping. Clinical concepts were extracted using ricu[23], an R package specifically designed to transform diverse datasets into a common semantically interoperable format. Following the extraction of a Sepsis Cohort and relevant variables, the pipeline facilitated data pre-processing and task definition, allowing us to extract harmonized cohorts from the three diverse databases. Moreover, the harmonized data structure comprised 48 dynamic features, including 7 vital signs, 39 laboratory tests, urine output, and FiO2. To ensure consistency across datasets, data frequency was harmonized to hourly intervals, as MIMIC-IV only provides hourly data input, while eICU and HiRID offer data entries at least every 5 minutes. The maximum duration of data collection was set at 7 days per patient.

2.1.1 Cohort Selection: Exclusion Criteria

For Sepsis Cohort selection, several exclusion criteria were applied in accordance with clinical definitions[22]. To ensure high data quality, we excluded patients with invalid admission or discharge times, ICU stays shorter than six hours, fewer than four hourly clinical measurements, or more than 12 consecutive hours without any clinical data. Additionally, we excluded patients under 18 years of age at admission and any stays from hospitals that did not report at least one sepsis case, to ensure sufficient data for defining the outcome. We also excluded cases where disease onset occurred outside the ICU or within the first six hours of ICU admission, enabling six-hour back-labeling of sepsis onset and supporting early prediction.

Dataset			HiRID	MIMIC-IV	eICU
Cohort Characteristics	Number of ICU stays		29,698	63,425	123,413
	Gender, n (%)	Male	19,037 (64.1)	35,170 (55.5)	66,888 (54.2%)
		Female	10,661 (35.9%)	28,255 (44.5%)	56,478 (45.8%)
	Age, n (%)	18-35 years	1,407 (4.7%)	4,777 (7.5%)	9,608 (7.8%)
		36-50 years	2,913 (9.8%)	7,844 (12.4%)	15,402 (12.5%)
		51-75 years	15,990 (53.8%)	33,131 (52.2%)	63,555 (51.5%)
		76-91 years	9,388 (31.6%)	15,310 (24.1%)	34,835 (28.2%)
	Body Mass Index (BMI) [kg/m^2]	$<$ 25 kg/m^2	12,621 (42.5%)	9,060 (14.3%)	37,968 (30.8%)
		25-29.9 kg/m^2	10,168 (34.2%)	9,046 (14.3%)	33,707 (27.3%)
		$\geq$ 30 kg/m^2	4,688 (15.8%)	9,308 (14.7%)	42,250 (34.2%)
		Unknown	2,221 (7.5%)	36,011 (56.8%)	9,488 (7.7%)
Sepsis Cases (Prevalence)	Total		1,858 (6.3%)	3,321 (5.2%)	5,638 (4.6%)
	Gender, n (%)	Male	1,272 (6.7%)	1,882 (5.4%)	3,099 (4.6%)
		Female	586 (5.5%)	1,439 (5.1%)	2,539 (4.5%)
	Age, n (%)	18-35 years	136 (9.7%)	212 (4.4%)	301 (3.1%)
		36-50 years	243 (8.3%)	388 (4.9%)	593 (3.9%)
		51-75 years	967 (6.0%)	1,770 (5.3%)	3,055 (4.8%)
		76-91 years	512 (5.5%)	839 (5.5%)	1,689 (4.8%)
	Body Mass Index (BMI) [kg/m^2]	$<$ 25 kg/m^2	882 (7.0%)	751 (8.3%)	1,877 (4.9%)
		25-29.9 kg/m^2	620 (6.1%)	701 (7.7%)	1,511 (4.5%)
		$\geq$ 30 kg/m^2	285 (6.1%)	794 (8.5%)	2,023 (4.8%)
		Unknown	71 (3.2%)	1,075 (3.0%)	227 (2.4%)

Table 2. Details of harmonized ICU cohorts HiRID, MIMIC-IV and eICU. Cohort characteristics and sepsis prevalence were analyzed following harmonization, with sepsis task and cohort definitions established according to the YAIB framework[22].

2.1.2 Sepsis Definition

Sepsis onset was defined as the first instance of organ dysfunction, characterized by a $\geq$2 point increase in the SOFA score within 24 hours, occurring within 48 hours before and 24 hours after suspected infection[17,29]. Suspected infection was identified based on the concurrent administration of antibiotics and the culture of body fluids, with cultures collected within 24 hours after antibiotic initiation or antibiotics started within 72 hours after culture collection. For HiRID and eICU, which lack microbiological data, suspicion of infection was determined solely by antibiotic administration. To exclude prophylactic use, antibiotics had to be administered for at least 3 days or for the entire ICU stay[30]. A "True" (Sepsis) label was assigned starting 6 hours before the determined onset, framing the task as a prediction problem, with a cutoff 12 hours after the onset, defining this as the onset period. This methodology yielded sepsis prevalences of 6.3% for HiRID, 5.2% for MIMIC-IV, and 4.6% for eICU (Table 2).

2.2 Data Preprocessing Pipeline

The preprocessing methodology followed a systematic approach to prepare intensive care unit data for machine learning analysis.

2.2.1 Clinical Domain-Informed Outlier Handling

Clinical measurements in intensive care settings are susceptible to errors arising from equipment malfunction or mishandling. The interpretation of vital signs and laboratory test results necessitates comprehensive evaluation of the clinical situation. Our harmonized datasets exhibited a considerable number of outliers across dynamic features. Each feature required careful evaluation to distinguish between dismissible (physiologically implausible) outliers and values indicative of severe patient conditions[31]. We implemented a comprehensive outlier detection approach grounded in clinical knowledge. For each physiological parameter, we defined both normal physiological ranges and extreme permissible values as given in Table 3.

These permissible limits were expanded by 5% of the permissible range in both directions and values outside these expanded ranges were converted to missing values (NaN) to preserve the temporal structure of the data.

The dataset was divided into training (90%) and testing (10%) sets using sklearn's GroupShuffleSplit, ensuring that all measurements from individual patient stays remained together. This approach prevented data leakage and maintained the temporal integrity of patient trajectories, which also taking into account admission time to the ICU, to obtain a temporal split. The resulting split preserved comparable frequencies of Sepsis Cases and Sepsis Labels between training and testing sets (Table 5).

We applied the Standard-Scaler from the Scikit-learn[32] library to transform both dynamic clinical measurements and static numeric features (age, weight, height). Each feature was standardized to zero mean and unit variance to facilitate model training, while sex as a categorical variable was encoded as -1/1. To prevent data leakage, mean and variance calculations were derived exclusively from the training set, and then applied to the testing set with learned mean and variance.

The handling of missing values proved particularly crucial given the nature of clinical time series data. Laboratory results frequently exhibited over 90% missing data, primarily due to daily measurement frequencies being represented in hourly data input. Comparing percentage and pattern of missing data across datasets and groups (sepsis, non-sepsis) did not reveal major differences that would prevent applying the same missing value strategy for all datasets and groups (Figure 2A, Figure 2B). We developed a two-pronged approach to address missing values. First, we created binary missingness indicators for each feature to explicitly capture missing value patterns. Second, we implemented hierarchical imputation, beginning with static features (demographics) using patient-specific forward/backward fill. For stay_ids lacking values, we assigned the global mean, and for gender, the mode (male). Dynamic features were imputed using a combination of forward fill within each patient stay and global mean imputation as a fallback, with global means calculated from the training set to prevent data leakage. This strategy was designed to be applicable in realistic clinical settings where new patients are admitted to the ICU.

2.3 Case-Control Matching

2.3.1 Matching Strategy and Criteria

Our overall preprocessing pipeline (see Figure 1A) begins with domain-informed outlier handling and proceeds through three distinct case-control balancing strategies: (Pathway 1) *undersampling* of random non-sepsis stays, (Pathway 2) *alignment* of randomly sampled non-sepsis stays using absolute onset matching for temporal window alignment, and (Pathway 3) *matching* with demographic criteria (*age*, *height*, *weight*, and *sex*) while also performing the same temporal alignment steps. These three strategies are then followed by data splitting, standardization, and missing value imputation, that were mentioned earlier. To illustrate the effect of these strategies, Figure 1C schematically contrasts absolute onset matching (with and without demographic variables) with undersampling, the latter lacking temporal alignment. To mitigate the class imbalance inherent in sepsis prediction, we employ the *matching* strategy (pathway 3 above) to ensure balanced representation of sepsis and non-sepsis cases while preserving clinical plausibility. In particular, *absolute onset matching* is used to prevent biases arising from systematically different length-of-stay distributions of sepsis vs. control groups[1,2]. Controls must have a monitoring duration at least as long as the matched case's pre-onset period, and data collected beyond 12 h post-matched absolute onset are excluded from further analysis. For demographic matching, primary matching criteria are: age within ± 5 years, height within ± 10 cm, weight within ± 10 kg, and exact matching for sex. Accounting for age and sex is standard practice in clinical studies[33,34], while weight and height (and thus approximate body-mass index) are included due to their reported effects on sepsis outcomes[35]. An ablation analysis was performed to investigate the contribution of these demographic features by removing them for both model training and testing. Across all three cohorts, the distribution of static features showed only minor variations across cases/controls and by gender (Figure 3), justifying a consistent approach.

2.3.2 Matching Algorithm Implementation

The matching process is summarized in Figure 1B ("matched", A3) and involves first identifying sepsis cases and their onset times. We then create a pool of non-sepsis controls that satisfy (i) the required demographic thresholds and (ii) the temporal alignment requirement (i.e., sufficient observation length to match each sepsis case). The *absolute* time since ICU admission is used to align each control's time series with the case's onset hour; a "time_to_onset" variable is introduced to track the relative

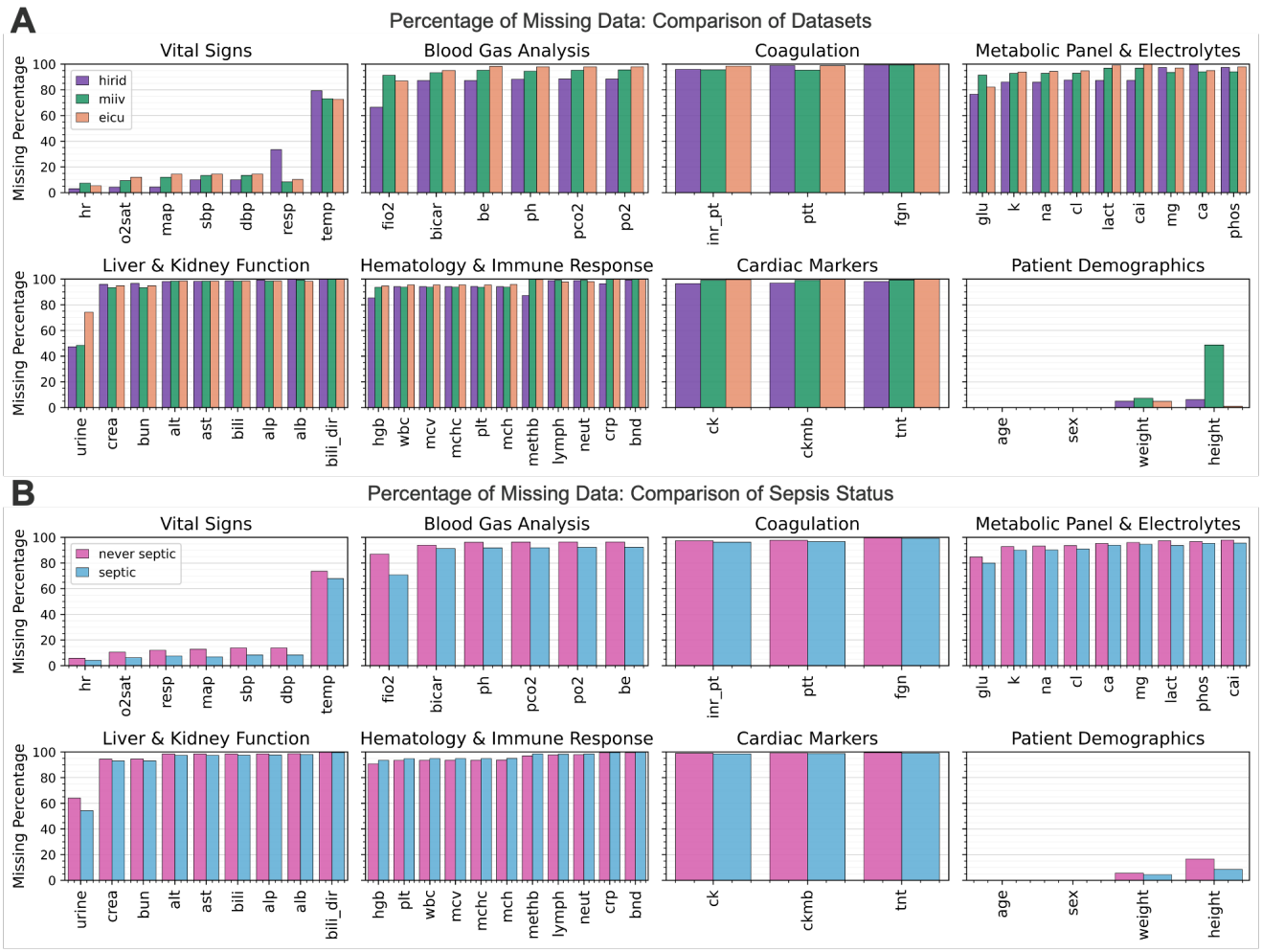


Figure 2. Percentage of missing data across features, datasets and sepsis status. Clinical time series data inherently exhibits substantial missingness due to the sparse and irregular sampling intervals employed during data collection. The multi-modal nature of intensive care unit datasets results in heterogeneous patterns of missing values across measurement types. Vital sign parameters are recorded at higher temporal frequencies compared to laboratory values, creating differential missingness profiles. Conversely, static patient characteristics such as demographic variables demonstrate near-complete data integrity with minimal missing entries. Percentages of missing data across all 48 dynamic and 4 static features in the harmonized datasets, stratified by dataset **(A)** and sepsis status (patients that are never septic and patients with confirmed or incipient sepsis) **(B)**.

position of each measurement. In instances where strict demographic matching cannot be met, we implement a two-tier fallback system to avoid losing sepsis cases[36]: (i) match on sex only, then (ii) allow any remaining controls with sufficient observation length. Each control is used only once (matching without replacement), thereby reducing selection bias. To investigate different effective sepsis prevalences, three target ratios of controls per sepsis case (10, 5, and 2) are sampled. Although this increases the proportion of sepsis in the final datasets relative to typical ICU prevalence, these values remain within ranges reported in multiple sepsis studies[2]. Matching was performed consistently all three cohorts, leading to decreasing overall cohort size as the sepsis ratio increased (Table 5, Figure 12). The same ratios were also applied during undersampling to provide a point of comparison, ensuring that performance increases or decreases are not merely due to changes in the number of patients within the cohorts, but rather the matching strategy itself.

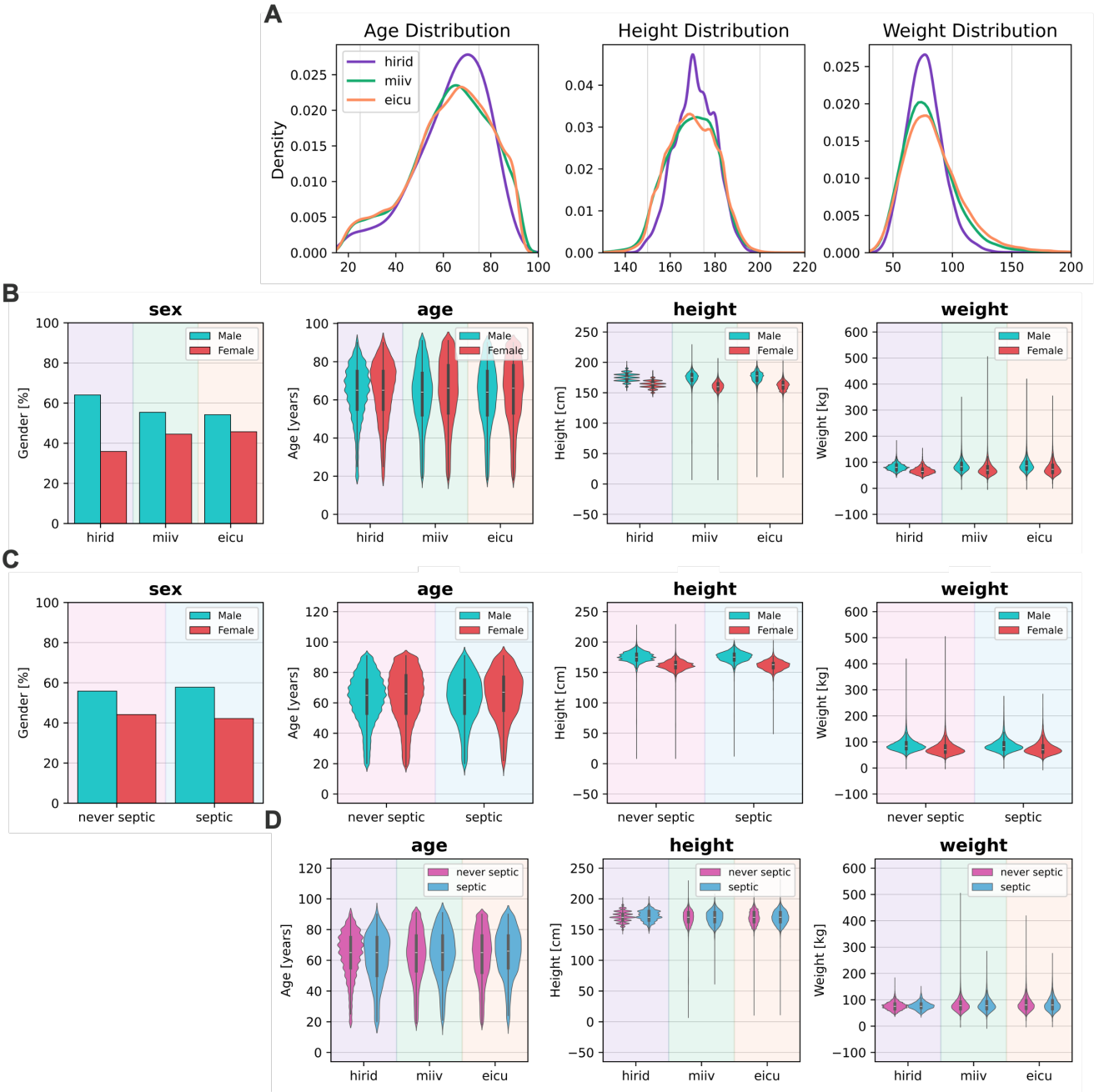


Figure 3. Distribution analysis of static features. A) Age, height, and weight distributions across three critical care datasets: HiRID, MIMIC-IV, and eICU. **B)** Comparative analysis of static clinical variables stratified by biological sex across the three datasets. **C)** Static clinical variables stratified by biological sex with comparative analysis between patients without sepsis (never septic) and those with confirmed or incipient sepsis (septic). **D)** Comparative analysis of static clinical variables stratified by age categories across the three critical care datasets.

2.4 Data Windowing

2.4.1 Window Design and Components

To transform continuous patient time series into structured input for machine learning models, we implemented a sliding window approach, sometimes referred to as "online training," which creates fixed-length data windows with varying prediction windows[2,30,37] (Figure 1D). Windowing was performed after preprocessing and potential case-control matching steps. This

methodology represents a standard approach for ICU prediction tasks based on structured time-series data, enabling models to identify temporal trends and patterns while evaluating their predictive capabilities across different time horizons. The sliding window technique offers greater sophistication than simply utilizing data from the minimum time interval (one hour) for sepsis classification. Notably, due to the six-hour backlabeling window applied during task definition, even this simpler implementation constitutes a six-hour prediction horizon[18,30,38,39]. We established a step length of 1 for the sliding window approach, maximizing the number of windows generated per stay_id. Only windows containing complete data and prediction windows were created, which resulted in fewer available windows with increasing window sizes.

The data window (observation window) was designed with a standardized length of 6 hours, which aligns with the YAIB exclusion criteria stipulating that patients must remain in the ICU for a minimum of six hours[22]. This parameter ensures that each stay_id generates at least one data window. The data window encompasses all dynamic clinical measurements and laboratory values while preserving the temporal sequence of observations. Static patient characteristics (age, sex, height, weight) were replicated across all windows to maintain consistency throughout the analysis. A gap between the data window and the targeted prediction time point is defined as a prediction window[40], while the prediction horizon comprises the sum of the prediction window and the backlabeling window. We implemented two prediction window configurations: first, no prediction window, utilizing the sepsis label from the final hour of the data window, resulting in a six-hour prediction horizon due to backlabeling; second, a four-hour prediction window, yielding a ten-hour prediction horizon to assess model performance for earlier prediction. This represents a prediction horizon that, while noted as challenging, has been demonstrated to be consistently achievable in previous research[41]. Earlier prediction capability would provide significant clinical benefit, allowing healthcare providers additional time to evaluate patient conditions and initiate treatment promptly. This consideration is particularly relevant as current machine learning models demonstrate higher accuracy in short-term rather than long-term sepsis prediction[42].

2.4.2 Implementation and Validation

The sliding window implementation processes each patient stay sequentially. For each possible starting position, a six-hour data window is created, with the values of all features being stacked (e.g., feature1_win0 until feature1_win5) and combined with the static patient information. The resulting data structure represents each window as a single row, containing both the stacked features and their corresponding label extracted from the appropriate prediction point. The validation process ensures data quality and temporal consistency through several critical requirements. Patient stays must be sufficiently long to accommodate both the data window and prediction window, with a step-length of 1 hour implemented to prevent data loss. The temporal integrity of measurements within each window is preserved, maintaining the original sequence and relationships of the clinical data. Furthermore, careful attention is paid to the alignment of outcome labels with their corresponding windows, ensuring accurate temporal relationships based on the designated prediction horizon. This comprehensive window creation methodology was applied uniformly across all three cohorts, as well as across all case-control matching variants (original, 1:10, 1:5, 1:2), ensuring consistency in the temporal representation of patient data throughout the analysis.

2.5 Linear Mixed Effects Models

2.5.1 Model Specification and Fitting

To assess the differences between septic and never-septic patients (cases and controls) while accounting for repeated measurements and individual patient variability, we implemented a Linear Mixed Effects Model (LMEM) analysis framework[43]. LMEMs incorporate both fixed effects, representing population-level parameters, and random effects, capturing variability within groups. For each clinical dynamic variable, we fitted a separate univariate LMEM of the form: $Variable \sim Group + (1 \mid PatientID)$ where the dependent variable represents the standardized clinical measurement (dynamic features), Group is a binary indicator (1 for septic cases, 0 for controls), and (1|PatientID) represents random intercepts for each patient stay. Models were fitted using restricted maximum likelihood estimation (REML)[44]. This model structure allows us to account for the repeated measurements of each variable within patients and estimate the fixed effect of sepsis status on each variable separately. Model diagnostics were performed, including assessment of convergence for each model, examination of variance components, and evaluation of model fit through random effects variance and residual analysis.

2.5.2 Statistical Analysis and Implementation

The statistical analysis framework incorporated multiple levels of inference and validation. P-values were adjusted for multiple comparisons using Benjamini-Hochberg FDR correction[45], and confidence intervals were calculated for fixed effects. Intraclass Correlation Coefficients (ICC)[46] were computed to assess the proportion of variance explained by patient-level clustering. The LMEM analysis was conducted across multiple dimensions, utilizing training sets that resulted from preprocessing with or without case-control matching and without windowing (Table 5, Figure 12). Models were fitted separately for each ICU dataset (HiRID, MIMIC-IV, eICU) to assess consistency across different data sources. The impact of case-control matching was evaluated by performing analyses on both the original and matched datasets, with multiple matching ratios (1:2, 1:5, 1:10) being considered. For effect size interpretation, coefficient estimates represent the standardized mean difference between cases and controls, with positive coefficients indicating higher values in septic patients. Effects were considered significant at an FDR-adjusted p-value < 0.05. This comprehensive approach allowed us to systematically evaluate the differences between septic and non-septic patients while accounting for the complex structure of intensive care unit data.

2.6 Machine Learning Models

2.6.1 Classification Models

For the prediction of sepsis onset, we implemented multiple machine learning approaches, encompassing both traditional ensemble methods and gradient boosting algorithms. Recent studies demonstrated equal or stronger performance of traditional machine learning models as compared to deep learning models for ICU prediction tasks[11–13]. Four distinct classification models were selected based on their specific characteristics relevant to sepsis prediction. Random Forest (RF)[47] and Balanced Random Forest (bRF)[48] were implemented using sklearn's RandomForestClassifier and imblearn's BalancedRandomForestClassifier respectively. While the standard RF implementation leverages ensemble learning with multiple decision trees maintaining the original class distribution, bRF incorporates automatic adjustments for class imbalance during training, making it particularly suitable for the natural class imbalance in sepsis prediction. Additionally, we employed two gradient boosting frameworks: LightGBM (LGBM)[49] and XGBoost (XGB)[50]. LGBM utilizes a leaf-wise tree growth strategy and was configured with a binary classification objective. XGBoost, incorporating additional regularization features, was optimized for binary classification tasks using AUC as the evaluation metric during training. To ensure reproducibility and establish baseline performance, all "vanilla" models were implemented with consistent configuration parameters, including a fixed random seed of 42, enabled parallel processing (n_jobs=-1), binary classification objectives where applicable, and default hyperparameters.

2.6.2 Model Optimization with Hyper-Parameter Tuning

For the best performing vanilla models, we optimized our model using a Bayesian optimization approach[51], utilizing a sequential approach to hyperparameter tuning that efficiently balances exploration and exploitation of the parameter space. The hyperparameter ranges for each model are specified in Table 6. The optimization protocol comprised five initial random points for exploration, followed by 25 sequential optimization iterations. Performance was evaluated using 5-fold cross-validation with AUROC as the optimization metric, and parallel processing was enabled. This optimization process was applied independently to each model and dataset combination, ensuring optimal adaptation to the specific characteristics of each ICU cohort while maintaining computational efficiency.

2.6.3 Evaluation of Model Performance

Model evaluation incorporated two complementary performance metrics: Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC). AUROC, while widely established in sepsis prediction literature and valuable for cross-study comparisons, exhibits limitations in imbalanced datasets due to its equal weighting of false positives and false negatives. An AUROC value of 0.5 represents performance equivalent to random classification, with values above 0.5 indicating increasingly better discrimination. To address the limitations of AUROC in class-imbalanced scenarios, we incorporated AUPRC as an additional evaluation metric, given its enhanced sensitivity to positive class predictions in imbalanced datasets. Since baseline Precision-Recall curves demonstrate dependence on class prevalence, we implemented normalization to enable valid cross-dataset comparisons. A normalized AUPRC (nAUPRC) of 1 reflected baseline performance

(random guessing), while values exceeding 1 indicated improved predictive performance compared to baseline. Normalized AUPRC was defined as:

$$nAUPRC = \frac{AUPRC}{Baseline}, \quad \text{where} \quad Baseline = \frac{Positive\ class\ samples}{Total\ samples}$$

The evaluation framework was designed to be dataset-agnostic, facilitating consistent comparison across the three ICU cohorts, while maintaining the ability to capture dataset-specific performance characteristics.

3 Results

3.1 Case-Control-Matching Reduces Significant Feature Differences Between Cases and Controls

Our LMEM analysis demonstrated that case-control matching substantially influenced FDR-corrected significance levels across features and datasets (Figure 4). A notable finding was the inverse relationship between sepsis prevalence and the retention of significant features, attributed to decreasing matching ratios (Figure 4A,B).

In the HiRID dataset, of the 48 time series variables analyzed, we observed a progressive reduction in significant features: 35 (100.0%) variables showed significant group differences in the original dataset, decreasing to 28 (80.0%) in the 1:10 matched training set, 25 (71.4%) in the 1:5 matched set, and 24 (68.6%) in the 1:2 matched set. The eICU dataset exhibited marginally more significant features initially, with 36 (100.0%) significant features in the original dataset, declining to 31 (86.1%) at 1:10 matching, 28 (77.8%) at 1:5 matching, and 24 (66.7%) at 1:2 matching. While the MIMIC-IV dataset presented the highest absolute number of significant features, it followed a similar declining pattern: 43 (100.0%) significant features in the original dataset, 35 (81.4%) at 1:10 matching, 32 (74.4%) at 1:5 matching, and 29 (67.4%) at 1:2 matching. These consistent patterns across all analyzed datasets suggest that case-control matching similarly affects the discriminative power of features between sepsis and non-sepsis patients.

Across ICU datasets, numerous dynamic features exhibited comparable significance levels (Figure 4C). We established significance for FDR-corrected p-values at the 0.05 threshold and further stratified significant features into three categories: $p<0.05$, $p<0.01$, and $p<0.001$. Vital signs including heart rate, systolic blood pressure, and mean arterial pressure demonstrated significant differences between sepsis and non-sepsis groups consistently across all datasets and matching variants. However, temperature showed reduced predictive value in HiRID, while oxygen saturation reached significance only in the original MIMIC-IV training set and across all matching variants of HiRID and eICU. The varying significance patterns across vital signs highlighted the complexity of sepsis detection using traditional clinical parameters.

Notably, while white blood cell count, a major inflammation marker, maintained significance throughout, C-reactive protein (CRP) showed consistent significance in HiRID variants but demonstrated reduced significance in MIMIC-IV and eICU. Lactate, a key indicator of tissue hypoperfusion and compromised oxygen delivery in septic shock, proved significant across all variants of HiRID and MIMIC-IV, but unexpectedly failed to reach significance in any eICU variant. These findings underscore the variability in the predictive value of established sepsis biomarkers across different clinical settings. Our analysis of organ function markers revealed distinct patterns, with urine output maintaining consistent significance while creatinine lost significance in matched variants of MIMIC-IV and eICU. Bilirubin, a surrogate marker for liver function, similarly showed reduced significance in certain variants of MIMIC-IV and eICU, suggesting potential regional differences in liver function assessment or documentation.

Despite dataset harmonization efforts, substantial differences in feature predictiveness persisted between the analyzed ICU cohorts. The most pronounced disparities were observed between HiRID, a Swiss dataset, and the American datasets MIMIC-IV and eICU (Table 1). These geographical variations in feature significance highlight the importance of considering regional healthcare practices and documentation standards when developing sepsis prediction models.

Our investigation next focused on quantifying the magnitude of LMEM coefficient changes across different matching ratio variants, both for aggregate features and individual parameters. In LMEMs with binary outcomes, coefficients indicate the probability change of outcome occurrence per unit increase in the predictor while accounting for random effects. A

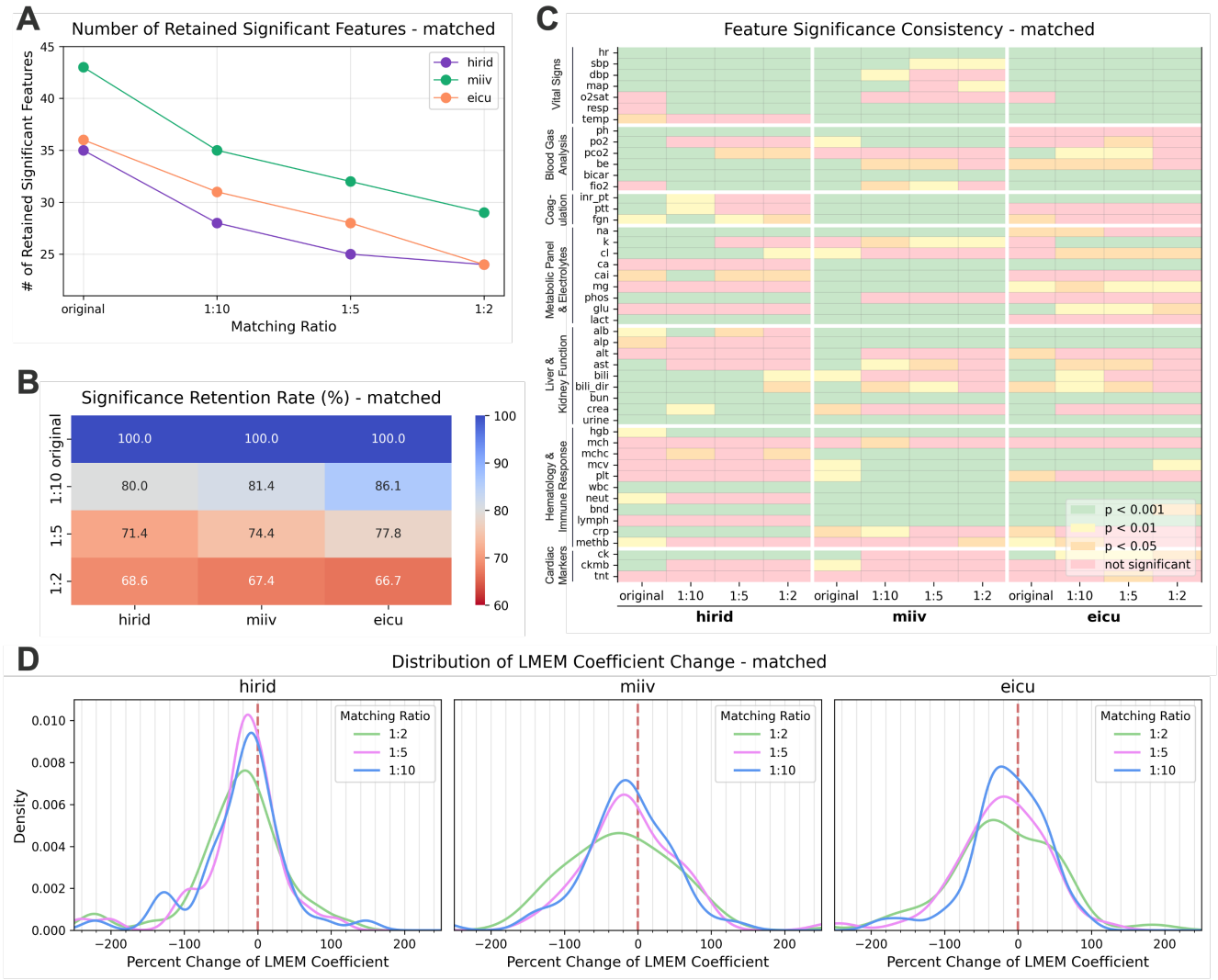


Figure 4. Evaluation of LMEM coefficient alterations and their statistical significance (original vs. matched). A) The absolute number of retained significant features in matched training cohorts compared to original training sets demonstrates an inverse relationship with increasing matching ratio. **B)** The relative significance retention rate exhibits comparable patterns across all three ICU datasets. **C)** Analysis of significance levels reveals distinct heterogeneity in feature significance profiles across datasets. **D)** The distribution shift toward negative percent changes in LMEM coefficients indicates reduced discriminative capacity of the collective dynamic features for differentiating between sepsis and non-sepsis cohorts.

coefficient approaching zero suggests a diminished expected effect on the outcome, which can be expressed as a negative Percent Change. The analysis revealed that matched training sets consistently demonstrated a shift toward negative Percent Changes across all ICU datasets and matching-ratio variants (Figure 4D). These findings were derived from the absolute values of LMEM coefficients (Figure 6), revealing distinct patterns across different clinical parameters. Dynamic features associated with vital signs, blood gas analysis, and metabolic panel and electrolytes demonstrated notably larger coefficient values compared to coagulation and cardiac markers. Interestingly, the 95% Confidence Intervals showed considerable variation across datasets, with eICU exhibiting the narrowest intervals and HiRID the widest, potentially attributable to differences in dataset size (Table 1). This variation in confidence intervals underscores the importance of considering dataset characteristics when interpreting the robustness of predictive features.

Moreover, the appendix (Figure 9 and Figure 10) shows the LMEM results comparing the original versus aligned cohorts (Pathway 2 in Figure 1A). The absolute number of retained significant features in aligned training cohorts, compared to the

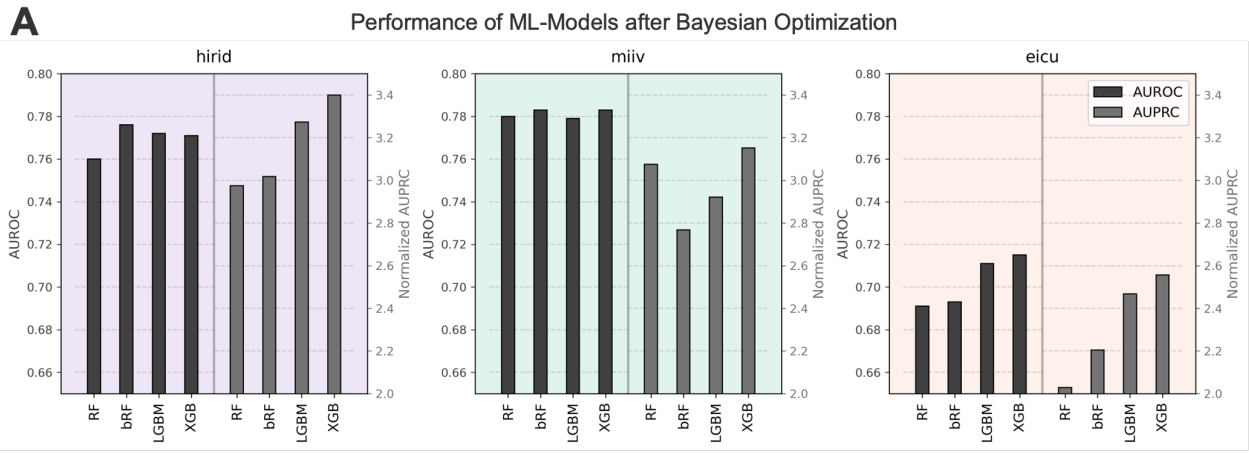
original training sets, displays an inverse relationship with increasing matching ratios. This relationship resembles the results from matched training cohorts (Pathway 3 in Figure 1A). The relative significance retention rate shows attenuated decrements with increasing matching ratios in MIMIC-IV compared to HiRID and eICU. LMEM coefficients mirror those derived from the LMEM analysis of matched training cohorts. The same analysis was also conducted to compare original and undersampled cohorts, in order to rule out the possibility that the above results are due solely to the reduced number of controls. Results in the appendix (Figure 7 and Figure 8) indicate that the absolute count of retained significant features in original versus undersampled training cohorts undergoes minimal reduction at 1:10 and 1:5 matching ratios. At the 1:2 matching ratio, dataset-specific patterns emerge: MIMIC-IV and, to a slightly lesser extent, eICU preserve nearly all significant features, whereas HiRID exhibits a substantial decrease. LMEM coefficients reveal negligible variation between original and undersampled training cohorts.

3.2 Sepsis Prediction Model Performance Declines After Case-Control Matching

The hyperparameter spaces utilized for Bayesian Optimization and their optimal values across different models and datasets are detailed in Table 6. XGBoost emerged as the slightly superior model across all three datasets when trained and tested with a one-hour data window and six-hour prediction horizon (see Methods section about windowing). The model achieved its highest performance on the HiRID dataset with an AUROC of 0.771 and nAUPRC of 3.400, followed by MIMIC-IV with an AUROC of 0.783 and nAUPRC of 3.152, while performance on eICU was lower with an AUROC of 0.715 and nAUPRC of 2.556 (Figure 5A). A consistent pattern emerged showing reduced model performance across all model architectures when trained and tested on the eICU dataset. Although the nAUPRC values demonstrated a 2-3.4-fold improvement over baseline predictive performance, the models still generated a number of false positive predictions, indicating room for further optimization.

We evaluated and compared model performance between undersampled (Pathway 1 in Figure 1A, case-control aligned (Pathway 2 in Figure 1A), case-control matched (Pathway 3 in Figure 1A), and original training sets without implementing model optimization strategies in Figure 5. All testing was conducted on the corresponding original testing sets to maintain consistency in evaluation. Our initial experiments utilized a one-hour observation window with a six hour prediction horizon (Figure 5B1), effectively requiring the models to predict labels for unmodified data rows. The models trained on original training sets demonstrated slightly lower AUROC and nAUPRC values compared to their Bayesian-optimized counterparts. Through ablation analysis, we discovered that excluding demographic features (sex, age, height, weight—shown as w/o demogr.) from both original training and testing sets did not negatively impact model performance. This finding further provides rationale for proceeding with our analysis of demographically case-control matched variants, as demographic features appeared non-essential for predictive accuracy. Substantial declines in model performance were observed across all datasets when training was conducted using case-control matched training sets (see Table 5 for training set sizes and sepsis prevalence). The severity of performance degradation was particularly striking, with AUROC values falling below 0.5, indicating performance worse than random chance, while nAUPRC values declined below 1, suggesting inferior performance compared to baseline predictions. These performance metrics demonstrated a clear inverse relationship with matching ratios, with the 1:2 matched cases and controls exhibiting the most pronounced deterioration in predictive capability. Interestingly, elevating sepsis prevalence through naive control undersampling significantly enhanced model performance, yielding maximum AUROC of 0.989 and maximum AUPRC of 41.09. A 1:10 matching ratio produced near-optimal AUROC metrics, with slight performance decrements observed at 1:5 and 1:2 ratios.Forest-based classifiers exhibited superior nAUPRC gains from undersampling (notably at 1:10 ratio) relative to gradient-boosted models. The same results are presented as AUPRC instead of nAUPRC in appendix (Figure 11). The appendix (Table 4) demonstrated comparable magnitudes of change in model performance metrics AUROC and AUPRC, with approximately half of the models showing slight improvements and half showing slight decrements (note that both AUROC and nAUPRC for these two matching methods are below/close to baseline performance and such small changes might not be meaningfully interpretable).

An interesting pattern emerged when comparing different model architectures: gradient-boosted models (LGBM and XGBoost) showed greater resilience in terms of AUROC values but demonstrated increased instability in nAUPRC metrics.



B — **Model Performance (vanilla)**

			AUROC										Normalized AUPRC									
			original	w/o demogr	undersampling			absolute onset matching			absolute onset matching + demographics criteria		original	w/o demogr	undersampling			absolute onset matching			absolute onset matching + demographics criteria	
					1:10	1:5	1:2	1:10	1:5	1:2	1:10	1:5	1:2		1:10	1:5	1:2	1:10	1:5	1:2	1:10	1:5	1:2
B1																							
1h data window, 6h prediction horizon	hirid	RF	0,727	0,727	0,988	0,976	0,978	0,461	0,345	0,357	0,405	0,346	0,319	2,542	2,417	27,94	14,27	6,883	0,513	0,362	0,274	0,419	0,265	0,206
		bRF	0,774	0,773	0,983	0,975	0,981	0,423	0,393	0,383	0,441	0,404	0,351	2,917	2,792	26,94	13,98	6,898	0,462	0,348	0,252	0,442	0,294	0,206
		LGBM	0,760	0,757	0,894	0,887	0,867	0,428	0,435	0,465	0,447	0,422	0,450	2,500	2,500	9,382	3,206	1,156	0,513	0,304	0,170	0,442	0,279	0,147
		XGB	0,720	0,732	0,915	0,917	0,894	0,422	0,439	0,481	0,442	0,430	0,446	2,583	2,708	12,24	5,365	1,469	0,462	0,333	0,193	0,442	0,294	0,147
	miiv	RF	0,711	0,705	0,989	0,990	0,970	0,372	0,320	0,308	0,389	0,370	0,331	2,231	1,208	27,62	15,05	6,750	0,256	0,145	0,141	0,233	0,221	0,154
		bRF	0,775	0,776	0,986	0,988	0,973	0,460	0,415	0,375	0,474	0,439	0,377	2,692	1,500	26,74	14,59	6,453	0,282	0,174	0,156	0,256	0,221	0,125
		LGBM	0,773	0,761	0,871	0,863	0,835	0,467	0,475	0,487	0,451	0,472	0,469	2,615	1,292	4,353	1,556	0,570	0,282	0,159	0,089	0,233	0,176	0,088
		XGB	0,755	0,743	0,908	0,901	0,867	0,462	0,458	0,455	0,460	0,442	0,472	2,615	1,458	7,912	2,698	0,766	0,282	0,159	0,081	0,256	0,176	0,088
	eicu	RF	0,619	0,629	0,975	0,966	0,969	0,367	0,328	0,332	0,397	0,338	0,337	1,545	0,750	26,65	14,05	6,961	0,256	0,159	0,119	0,233	0,176	0,176
		bRF	0,692	0,688	0,969	0,967	0,970	0,471	0,435	0,383	0,477	0,426	0,383	2,273	1,000	25,82	13,51	6,758	0,308	0,217	0,133	0,256	0,191	0,191
		LGBM	0,697	0,703	0,794	0,784	0,779	0,494	0,493	0,483	0,494	0,475	0,490	2,273	1,083	2,500	0,937	0,367	0,282	0,159	0,081	0,256	0,147	0,088
		XGB	0,675	0,677	0,846	0,835	0,816	0,507	0,488	0,483	0,496	0,478	0,485	2,091	1,000	3,500	1,683	0,445	0,308	0,159	0,081	0,256	0,162	0,081
B2																							
6h data window, 6h prediction horizon	hirid	RF	0,735	0,733	0,989	0,973	0,978	0,483	0,367	0,378	0,425	0,405	0,374	5,308	5,308	41,09	21,31	9,898	0,719	0,421	0,351	0,594	0,368	0,281
		bRF	0,784	0,781	0,985	0,974	0,979	0,468	0,409	0,371	0,464	0,431	0,398	6,000	6,077	40,04	20,50	9,227	0,688	0,421	0,316	0,656	0,404	0,316
		LGBM	0,785	0,773	0,914	0,906	0,889	0,452	0,429	0,442	0,478	0,475	0,460	6,615	6,231	17,35	6,833	2,318	0,656	0,439	0,228	0,719	0,404	0,193
		XGB	0,754	0,760	0,934	0,931	0,932	0,423	0,409	0,450	0,478	0,462	0,435	6,615	6,308	21,65	9,881	3,659	0,594	0,351	0,237	0,688	0,368	0,202
	miiv	RF	0,763	0,763	0,991	0,990	0,972	0,403	0,346	0,345	0,428	0,400	0,350	3,308	3,385	40,96	22,48	9,614	0,313	0,175	0,175	0,313	0,263	0,184
		bRF	0,812	0,810	0,987	0,990	0,972	0,493	0,460	0,400	0,487	0,454	0,422	3,769	3,846	38,61	20,29	7,830	0,375	0,211	0,184	0,375	0,263	0,167
		LGBM	0,817	0,816	0,901	0,910	0,883	0,474	0,477	0,488	0,448	0,434	0,475	3,692	3,692	8,261	3,405	1,114	0,344	0,193	0,096	0,313	0,193	0,105
		XGB	0,802	0,803	0,935	0,926	0,901	0,467	0,458	0,470	0,455	0,450	0,460	4,000	4,231	13,83	5,524	1,682	0,344	0,175	0,096	0,344	0,211	0,105
	eicu	RF	0,657	0,655	0,975	0,970	0,972	0,397	0,367	0,368	0,426	0,367	0,362	1,692	1,769	39,35	21,07	10,03	0,313	0,211	0,149	0,313	0,246	0,211
		bRF	0,730	0,730	0,968	0,970	0,971	0,527	0,482	0,424	0,520	0,469	0,435	2,538	2,538	36,30	18,60	8,341	0,438	0,281	0,175	0,406	0,281	0,237
		LGBM	0,740	0,733	0,820	0,820	0,812	0,512	0,507	0,481	0,498	0,485	0,498	2,846	2,692	2,696	1,714	0,761	0,375	0,211	0,096	0,344	0,193	0,105
		XGB	0,721	0,716	0,876	0,866	0,860	0,510	0,499	0,499	0,493	0,472	0,491	2,615	2,615	7,565	2,929	1,068	0,375	0,211	0,105	0,344	0,193	0,105
B3																							
6h data window, 10h prediction horizon	hirid	RF	0,716	0,706	0,986	0,975	0,976	0,479	0,339	0,361	0,419	0,398	0,385	5,818	5,909	40,91	20,81	9,798	0,733	0,364	0,404	0,633	0,389	0,287
		bRF	0,762	0,759	0,982	0,971	0,974	0,473	0,425	0,391	0,474	0,433	0,420	6,818	6,455	39,61	19,26	8,112	0,767	0,400	0,339	0,700	0,481	0,306
		LGBM	0,773	0,773	0,926	0,913	0,901	0,446	0,440	0,449	0,478	0,446	0,482	6,636	7,545	16,65	6,302	2,315	0,633	0,436	0,193	0,733	0,370	0,213
		XGB	0,723	0,738	0,952	0,934	0,928	0,440	0,436	0,483	0,448	0,464	0,484	7,727	7,727	25,48	9,349	3,506	0,633	0,418	0,284	0,667	0,407	0,250
	miiv	RF	0,732	0,733	0,990	0,988	0,966	0,393	0,340	0,336	0,415	0,405	0,340	3,364	3,091	40,87	21,79	9,427	0,300	0,200	0,183	0,300	0,241	0,167
		bRF	0,783	0,781	0,985	0,988	0,961	0,509	0,446	0,413	0,489	0,452	0,405	3,818	3,818	37,22	18,05	5,831	0,400	0,200	0,202	0,400	0,259	0,148
		LGBM	0,785	0,773	0,903	0,907	0,881	0,479	0,499	0,497	0,449	0,458	0,486	3,364	3,273	7,478	3,023	1,034	0,333	0,218	0,092	0,300	0,185	0,102
		XGB	0,759	0,747	0,938	0,944	0,900	0,472	0,468	0,485	0,479	0,460	0,480	3,545	3,636	14,70	5,977	1,258	0,333	0,200	0,101	0,333	0,185	0,102
	eicu	RF	0,646	0,645	0,971	0,970	0,970	0,388	0,366	0,350	0,427	0,354	0,357	1,909	1,909	39,39	20,63	9,910	0,333	0,218	0,147	0,333	0,222	0,213
		bRF	0,709	0,709	0,964	0,968	0,967	0,519	0,492	0,433	0,531	0,476	0,442	2,818	2,727	35,35	16,79	6,809	0,467	0,309	0,193	0,433	0,315	0,250
		LGBM	0,702	0,707	0,813	0,817	0,815	0,520	0,511	0,488	0,519	0,495	0,491	2,636	2,727	3,609	1,674	0,685	0,400	0,200	0,101	0,400	0,204	0,102
		XGB	0,694	0,700	0,884	0,875	0,857	0,519	0,507	0,488	0,502	0,504	0,490	2,727	2,818	7,348	3,070	1,045	0,367	0,236	0,101	0,367	0,222	0,102

0,31 — 0,99 (AUROC)

0,08 — 41,1 (Normalized AUPRC)

Figure 5. Performance of machine learning models. A) AUROC and nAUPRC metrics of machine learning models following Bayesian hyperparameter optimization, trained and evaluated on original datasets. **B)** Comparative analysis of non-optimized (vanilla) machine learning models trained on specified training datasets (original, demographics-excluded, as well as undersampled, aligned, and matched at 1:10/1:5/1:2 ratios each) and evaluated on the original test cohort. Left panels display AUROC metrics while right panels present nAUPRC metrics. This experimental framework was implemented across three distinct temporal configurations: **(B1)** 1-hour data window with 0-hour prediction window, **(B2)** 6-hour data window with 0-hour prediction window, and **(B3)** 6-hour data window with 4-hour prediction window. Models trained on undersampled cohorts demonstrate substantial performance metric improvements, while those trained on absolute-onset matched cohorts—with or without demographic criteria—exhibit performance metrics below baseline levels.

Analysis of relative performance revealed a noteworthy cohort-specific effect, with the eICU cohort exhibiting more modest performance degradation compared to both HiRID and MIMIC-IV cohorts. This variable impact of case-control matching across cohorts suggests underlying differences in data structure and characteristics that influence model robustness. Examining different data and prediction window configurations revealed consistent patterns across our experiments. Performance peaked when models utilized six hours of input data with six hours of prediction horizon (Figure 5B2), demonstrating enhanced sepsis prediction capability under these conditions. The MIMIC-IV dataset yielded particularly decent results, with the LGBM model achieving the highest AUROC value of 0.817 and the XGBoost model reaching a peak nAUPRC value of 3.998. This may be due to the exceeding absolute numbers of significant group differences, as shown in the LMEM analysis (Figure 4B), paired with the data origin from a single hospital system. When we extended the prediction horizon to 10 hours beyond the data window while maintaining the six-hour data window, model performance aligned closely with our baseline models that used an one-hour data window and six-hour prediction horizon (Figure 5B3). A consistent trend emerged across all three windowing variants: models trained and tested on HiRID or MIMIC-IV consistently outperformed those utilizing the eICU dataset.

In summary, our experimental framework yielded model performance metrics that align with contemporary research findings across all investigated datasets, model algorithms, and temporal window configurations when using the original training set structure. However, when implementing absolute onset case-control matching with and without demographic characteristics to the training sets (both Pathway 2 and Pathway 3 in Figure 1A), we observed a substantial decrease in model performance across all experimental parameters. The matching ratio exhibited only a minor influence on performance outcomes, with an inverse correlation between model performance and matching ratio values. This suggests a paradoxical relationship in which absolute onset case-control matched training sets exhibit progressively higher sepsis prevalence, yet models fail to benefit from the reduced class imbalance and instead demonstrate difficulty in identifying predictive patterns.

4 Discussion

The goal of sepsis prediction algorithms is to identify patients at risk before the infection reaches a stage where treatment is less effective or major complications arise[52]. This remains a challenging task, as machine learning classification performance for sepsis demonstrates markedly lower accuracy compared to other ICU prediction tasks, such as mortality or acute kidney injury prognosis[53,54]. Enhanced methodological approaches are required to address the inherent challenges of pronounced class imbalance and optimize model performance. Clinical implementation necessitates robust model performance with an optimal balance between sensitivity and specificity to establish credibility among healthcare professionals.

4.1 The Matching Paradox in Sepsis Prediction

Many previous sepsis prediction studies have not explicitly used case-control matching, instead relying on observational datasets with imbalanced class distributions[30,37,55,56]. This methodological choice may contribute to the reported strong model performance, raising a critical paradox: could the absence of structured matching itself be influencing the observed predictive success? By addressing this issue, our study challenges the community to reconsider the best approach for developing safe and unbiased predictive models for sepsis prediction. We argue that understanding and mitigating potential biases arising from dataset construction choices is crucial for ensuring robust and generalizable clinical AI applications.

Our LMEM analysis revealed that case-control matching systematically reduced the number of features showing significant differentiation between sepsis and non-sepsis cases across all three ICU datasets. From an initial set of 35–43 significant features, matching at a 1:2 ratio retained only 24–29 significant features, representing a reduction of approximately 30%. While vital signs such as heart rate and blood pressure maintained significance across all matching variants, other traditionally important sepsis indicators displayed variable robustness. Notably, inflammatory markers like CRP exhibited reduced significance in MIMIC-IV and eICU matched variants, while lactate, a key indicator of tissue hypoperfusion, lost significance in the eICU variants. Moreover, we also found that performing "alignment" instead of full-fledged matching led to similar performance deteriorations. The higher performance observed in undersampled cohorts, compared to the original dataset, confirms that performance deterioration is attributable to the matching criteria rather than merely the reduced sample size. These findings suggest that matching procedures may inadvertently dilute crucial biological signals that serve as early sepsis indicators.

Our analysis revealed an inverse relationship between matching stringency and prediction model performance across all three harmonized ICU cohorts and the evaluated model algorithms, challenging conventional epidemiological assumptions about the benefits of matching for inference. The most severe performance degradation occurred with 1:2 matching, where AUROC values dropped below 0.5 and nAUPRC values fell below baseline levels. Notably, this performance drop was not simply due to data loss from matching—undersampled cohorts with the same ratio showed higher performance than the original unbalanced cohorts. Thus, the observed deterioration (across both alignment- and demographic-based matching) persisted despite increased sepsis prevalence in the matched datasets, indicating that matching-induced distortions in the feature space outweighed any potential benefits from class rebalancing. These findings were consistent across different temporal windows, with models showing optimal performance when using six hours of input data and minimal prediction windows.

Two potential mechanisms could explain these observations: (1) case-control matching may create artificial data distributions that obscure naturally occurring sepsis signals, particularly affecting gradient-boosted models, which showed varying stability in AUROC versus nAUPRC metrics; and (2) machine learning models developed without case-control matching may learn temporal patterns that, while predictive, are not directly related to sepsis pathophysiology. Our ablation analysis revealed that demographic features contributed minimally to prediction performance in unmatched datasets, suggesting these variables may not be critical for short-term sepsis prediction. However, recent research demonstrated that feature importance can vary significantly across prediction windows, with gender contributing 22% to 24-hour predictions but only 3% to 6-hour forecasts[57]. These findings highlight fundamental challenges in balancing epidemiological best practices with machine learning approaches for sepsis prediction, particularly when working across diverse clinical datasets and healthcare systems, and underscore the importance of considering prediction horizons when evaluating feature contributions.

4.2 Future Directions of Case-Control Matching

Several approaches could potentially mitigate the limitations observed with traditional demographic case-control matching in sepsis prediction. One promising direction involves expanding matching criteria beyond demographics to include baseline clinical characteristics. Implementing pre-onset trajectory matching, in which controls are matched based on vital sign trajectories during early ICU admission, could better account for temporal dynamics[58]. Similarly, creating strata using time-conditional percentiles of dynamic biomarkers such as lactate or white blood cell counts might improve matching precision. However, these approaches face practical constraints, especially related to data availability prior to sepsis onset; many patients meet sepsis criteria shortly after ICU admission, and blood-based clinical features often have delayed availability. Moreover, propensity score matching presents an alternative solution to address the dimensionality challenges inherent in traditional case-control matching[59]. This method calculates a composite score representing each patient's probability of belonging to the case group based on multiple covariates, allowing for more efficient group balancing and enabling sophisticated matching approaches such as nearest neighbor or caliper matching. Although propensity score methods offer flexibility and reduce the need for arbitrary categorization of continuous variables, they generally require larger sample sizes for accurate score estimation.

Machine learning methods also offer new possibilities for dynamic matching optimization during model training. For instance, reinforcement learning-based feature masking could selectively retain features with high sepsis discrimination power while maintaining control over confounding variables. The DeepAISE framework illustrates this potential, achieving AUROCs between 0.87 and 0.90 for predictions made 4–12 h prior to sepsis onset through dynamic feature weighting across 24-hour windows[60]. These findings suggest that similar reinforcement architectures could optimize matching without compromising predictive capability. Additionally, generative adversarial networks (GANs) could synthesize matched controls, creating balanced datasets that preserve temporal covariate relationships while retaining critical sepsis signals. However, although synthetic controls offer advantages in data availability and near-perfect matching capabilities, they may introduce artificial patterns and lack the inherent biological variability found in real patient data. Consequently, the generalizability of models trained on such synthetic datasets could be limited.

4.3 Consequences of Dataset Harmonization

Recent advances in dataset harmonization across major publicly available ICU databases, such as HiRID, MIMIC-IV, and eICU, have substantially enhanced opportunities for multi-center validation of predictive models in intensive care research. However, while these harmonization processes standardize variable definitions and temporal resolutions, they may inadvertently compromise predictive power. Performance metrics from YAIB experiments[22] demonstrate lower accuracy compared to state-of-the-art models trained on raw institutional data and similar publicly available datasets[2,56]. This discrepancy partly reflects YAIB's focus beyond pure performance optimization, yet several specific harmonization-related impacts on data structure and granularity contribute to this performance gap. The standardization process can discard institution-specific predictive features embedded in local treatment protocols and documentation practices, effectively treating these latent prognostic signals as statistical noise. Furthermore, harmonized exclusion criteria can result in artificially homogeneous patient populations that lack the clinical heterogeneity necessary for developing robust, generalizable models. The temporal alignment of irregular clinical measurements to standardized time grids particularly affects the preservation of natural physiological trajectories crucial for dynamic prediction. This impact is especially pronounced for HiRID and eICU, which originally capture data at higher temporal resolutions than MIMIC-IV's hourly intervals, leading to a significant loss of granular temporal information during harmonization. Nevertheless, these factors do not negate the fact that alignment and matching significantly diminished predictive power in comparison with unbalanced or original cohorts. It is also unwarranted to assume that even using non-harmonized versions of the datasets would have produced different results.

In terms of study limitations, several factors affect the interpretation and generalizability of our findings. Traditional case-control matching introduces intrinsic dimensionality challenges when matching characteristics are considered separately, particularly as the need for non-reusable matches increases the complexity of matching. While alternative approaches such as sex-based or random matching can help mitigate match availability issues, these simplified strategies may lead to suboptimal confounder control across the entire cohort. Our study's cohort selection criteria and sepsis definition, following the YAIB workflow, resulted in notably lower sepsis prevalence compared to previous research using the same datasets. For example, a recent study reported a sepsis prevalence of 37% in HiRID[61], while our analysis identified only 6.3%. The use of harmonized datasets further constrained feature availability. Although the SOFA score plays a central role in defining sepsis onset, critical components like vasopressor administration and Glasgow Coma Scale (GCS) scores were not available in these harmonized dataset versions. Additional clinician-guided harmonization efforts will be essential to expand the feature set, particularly regarding the medications and comorbidities[22]. Another limitation stems from our temporal definition of sepsis onset, which encompassed Sepsis-3 defined onset within a window spanning 48 h prior to and 24 h after the clinical suspicion of sepsis. This broad window complicates the ability to determine whether our models could predict sepsis earlier than the time of clinical suspicion. As noted in recent literature[52], machine learning predictors should ideally identify at-risk patients before clinical suspicion arises—an ability our current study design cannot definitively demonstrate.

4.4 Conclusion

Our investigation highlights a fundamental challenge in sepsis prediction: models trained on unmatched data may capture patterns reflecting general ICU trajectories rather than sepsis-specific pathophysiology, potentially due to uncontrolled variations in length of stay and patient-specific monitoring durations. While case-control matching presents a promising strategy for improving model specificity and robustness, our results emphasize that matching methodologies must be carefully designed to preserve clinically relevant sepsis signals. The significant performance degradation observed with increasing matching stringency suggests that overly rigid approaches may inadvertently discard crucial predictive features. Striking the right balance between confounder control and preserving sepsis-specific patterns remains a critical challenge. Our findings underscore the need for more sophisticated preprocessing strategies in developing reliable machine learning models for sepsis prediction. Future research should focus on refining matching techniques to mitigate temporal bias while retaining pathophysiologically relevant signals—potentially by integrating dynamic matching constraints and clinical domain knowledge. Ultimately, by combining epidemiological rigor with advanced machine learning techniques, the field can move closer to delivering clinically actionable, generalizable sepsis detection tools that address the real-world complexities of critical care.

References

1. Moor, M., Horn, M., Rieck, B., Roqueiro, D. & Borgwardt, K. Early Recognition of Sepsis with Gaussian Process Temporal Convolutional Networks and Dynamic Time Warping. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, 2–26 (PMLR, 2019). ISSN: 2640-3498.

2. Moor, M., Rieck, B., Horn, M., Jutzeler, C. R. & Borgwardt, K. Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review. *Front. Medicine* **8**, DOI: 10.3389/fmed.2021.607952 (2021). Publisher: Frontiers.

3. Futoma, J. *et al.* An Improved Multi-Output Gaussian Process RNN with Real-Time Validation for Early Sepsis Detection. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, 243–254 (PMLR, 2017). ISSN: 2640-3498.

4. Futoma, J., Hariharan, S. & Heller, K. Learning to Detect Sepsis with a Multitask Gaussian Process RNN Classifier. In *Proceedings of the 34th International Conference on Machine Learning*, 1174–1182 (PMLR, 2017). ISSN: 2640-3498.

5. Bauer, M. *et al.* Mortality in sepsis and septic shock in Europe, North America and Australia between 2009 and 2019- results from a systematic review and meta-analysis. *Critical Care (London, England)* **24**, 239, DOI: 10.1186/s13054-020-02950-2 (2020).

6. Evans, L. *et al.* Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Intensive Care Medicine* **47**, 1181–1247, DOI: 10.1007/s00134-021-06506-y (2021).

7. De Silva, M., Chadwick, W. & Naidoo, N. Screening tools for sepsis identification in paramedicine and other emergency contexts: a rapid systematic review. *Scand. journal trauma, resuscitation emergency medicine* **31**, 74 (2023).

8. Ruppel, H. & Liu, V. To catch a killer: electronic sepsis alert tools reaching a fever pitch? (2019).

9. Bomrah, S. *et al.* A scoping review of machine learning for sepsis prediction- feature engineering strategies and model performance: a step towards explainability. *Critical Care* **28**, 180, DOI: 10.1186/s13054-024-04948-6 (2024).

10. Hüser, M. *et al.* RMS: A ML-based system for ICU Respiratory Monitoring and Resource Planning, DOI: 10.1101/2024.01.23.24301516 (2024). Pages: 2024.01.23.24301516.

11. Burger, M. *et al.* Towards Foundation Models for Critical Care Time Series, DOI: 10.48550/arXiv.2411.16346 (2024). ArXiv:2411.16346 [cs].

12. Hyland, S. L. *et al.* Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat. Medicine* **26**, 364–373, DOI: 10.1038/s41591-020-0789-4 (2020). Publisher: Nature Publishing Group.

13. Lyu, X. *et al.* An empirical study on KDIGO-defined acute kidney injury prediction in the intensive care unit. *Bioinforma. (Oxford, England)* **40**, i247–i256, DOI: 10.1093/bioinformatics/btae212 (2024).

14. Goh, K. H. *et al.* Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nat. communications* **12**, 711 (2021).

15. Islam, K. R. *et al.* Machine learning-based early prediction of sepsis using electronic health records: a systematic review. *J. clinical medicine* **12**, 5658 (2023).

16. Wong, A. *et al.* External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA internal medicine* **181**, 1065–1070 (2021).

17. Singer, M. *et al.* The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* **315**, 801, DOI: 10.1001/jama.2016.0287 (2016).

18. Do, D.-K. *et al.* The impact of evaluation strategy on sepsis prediction model performance metrics in intensive care data, DOI: 10.1101/2025.02.20.25322509 (2025). Pages: 2025.02.20.25322509.

19. Iwagami, M. & Shinozaki, T. Introduction to matching in case-control and cohort studies. *Annals Clin. Epidemiol.* **4**, 33–40 (2022).

20. Wan, F., Colditz, G. A. & Sutcliffe, S. Matched versus unmatched analysis of matched case-control studies. *Am. journal epidemiology* **190**, 1859–1866 (2021).

21. Mansournia, M. A., Jewell, N. P. & Greenland, S. Case–control matching: effects, misconceptions, and recommendations. *Eur. journal epidemiology* **33**, 5–14, DOI: 10.1007/s10654-017-0325-0 (2018).

22. Water, R. v. d. *et al.* Yet Another ICU Benchmark: A Flexible Multi-Center Framework for Clinical ML, DOI: 10.48550/arXiv.2306.05109 (2024). ArXiv:2306.05109 [cs].

23. Bennett, N., Plečko, D., Ukor, I.-F., Meinshausen, N. & Bühlmann, P. ricu: R's interface to intensive care data. *GigaScience* **12**, giad041, DOI: 10.1093/gigascience/giad041 (2023). Accepted: 2023-08-24T10:06:07Z Publisher: Oxford University Press.

24. Faltys, M. *et al.* HiRID, a high time-resolution ICU dataset, DOI: 10.13026/MN2D-NQ25.

25. Johnson, A. E. W. *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1, DOI: 10.1038/s41597-022-01899-x (2023). Publisher: Nature Publishing Group.

26. Johnson, A. *et al.* MIMIC-IV, DOI: 10.13026/KPB9-MT58.

27. Pollard, T. J. *et al.* The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci. Data* **5**, 180178, DOI: 10.1038/sdata.2018.178 (2018).

28. Pollard, T. J., Johnson, A. E. W., Raffa, J. & Badawi, O. The eICU Collaborative Research Database, DOI: 10.13026/C2WM1R (2017).

29. Seymour, C. W. *et al.* Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* **315**, 762–774, DOI: 10.1001/jama.2016.0288 (2016).

30. Reyna, M. *et al.* Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019. DOI: 10.22489/CinC.2019.412 (2019).

31. Shamout, F., Zhu, T. & Clifton, D. A. Machine Learning for Clinical Outcome Prediction. *IEEE Rev. Biomed. Eng.* **14**, 116–126, DOI: 10.1109/RBME.2020.3007816 (2021). Conference Name: IEEE Reviews in Biomedical Engineering.

32. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

33. Iwagami, M. & Shinozaki, T. Introduction to Matching in Case-Control and Cohort Studies. *Annals Clin. Epidemiol.* **4**, 33–40, DOI: 10.37737/ace.22005 (2022).

34. Lauritsen, S. M. *et al.* Early detection of sepsis utilizing deep learning on electronic health record event sequences. *Artif. Intell. Medicine* **104**, 101820, DOI: 10.1016/j.artmed.2020.101820 (2020).

35. Bai, L. *et al.* Association of body mass index with mortality of sepsis or septic shock: an updated meta-analysis. *J. Intensive Care* **11**, 27, DOI: 10.1186/s40560-023-00677-0 (2023).

36. Ratanatharathorn, A., Mooney, S. J., Rybicki, B. A. & Rundle, A. G. A Flexible Matching Strategy for Matched Nested Case Control Studies. *Annals epidemiology* **86**, 49–56.e3, DOI: 10.1016/j.annepidem.2023.06.023 (2023).

37. Kok, C. *et al.* Automated prediction of sepsis using temporal convolutional network. *Comput. Biol. Medicine* **127**, 103957, DOI: 10.1016/j.compbiomed.2020.103957 (2020).

38. Ewig, K. *et al.* Multi-Subset Approach to Early Sepsis Prediction, DOI: 10.48550/arXiv.2304.06384 (2023). ArXiv:2304.06384 [cs].

39. Moor, M. *et al.* Predicting sepsis in multi-site, multi-national intensive care cohorts using deep learning, DOI: 10.48550/arXiv.2107.05230 (2021). ArXiv:2107.05230 [cs].

40. Bloch, E., Rotem, T., Cohen, J., Singer, P. & Aperstein, Y. Machine Learning Models for Analysis of Vital Signs Dynamics: A Case for Sepsis Onset Prediction. *J. Healthc. Eng.* **2019**, 5930379, DOI: 10.1155/2019/5930379 (2019). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2019/5930379.

41. Valik, J. K. *et al.* Predicting sepsis onset using a machine learned causal probabilistic network algorithm based on electronic health records data. *Sci. Reports* **13**, 11760, DOI: 10.1038/s41598-023-38858-4 (2023). Publisher: Nature Publishing Group.

42. Yadgarov, M. Y. *et al.* Early detection of sepsis using machine learning algorithms: a systematic review and network meta-analysis. *Front. Medicine* **11**, DOI: 10.3389/fmed.2024.1491358 (2024). Publisher: Frontiers.

43. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48, DOI: 10.18637/jss.v067.i01 (2015).

44. Patterson, H. D. & Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554 (1971).

45. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).

46. Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. chiropractic medicine* **15**, 155–163 (2016).

47. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32, DOI: 10.1023/A:1010933404324 (2001).

48. Chen, C., Liaw, A. & Breiman, L. Using random forest to learn imbalanced data. *Univ. California, Berkeley* **110**, 1–12 (2004).

49. Zhang, J., Mucs, D., Norinder, U. & Svensson, F. Lightgbm: An effective and scalable algorithm for prediction of chemical toxicity–application to the tox21 and mutagenicity data sets. *J. chemical information modeling* **59**, 4150–4158 (2019).

50. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (2016).

51. Snoek, J., Larochelle, H. & Adams, R. P. Practical bayesian optimization of machine learning algorithms. *Adv. neural information processing systems* **25** (2012).

52. Logan Ellis, H. *et al.* The early warning paradox. *npj Digit. Medicine* **8**, 1–2, DOI: 10.1038/s41746-024-01408-x (2025). Publisher: Nature Publishing Group.

53. Lu, S.-C. *et al.* Machine Learning-Based Short-Term Mortality Prediction Models for Patients With Cancer Using Electronic Health Record Data: Systematic Review and Critical Appraisal. *JMIR medical informatics* **10**, e33182, DOI: 10.2196/33182 (2022).

54. Huang, W., Chen, Y., Wang, P., Liu, X. & Liu, S. An Interpretable Temporal Convolutional Network Model for Acute Kidney Injury Prediction in the Intensive Care Unit. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 3021–3028, DOI: 10.1109/BIBM52615.2021.9669653 (2021).

55. Merath, K. *et al.* Use of Machine Learning for Prediction of Patient Risk of Postoperative Complications After Liver, Pancreatic, and Colorectal Surgery. *J. Gastrointest. Surg.* **24**, 1843–1851, DOI: 10.1007/s11605-019-04338-2 (2020).

56. Wang, Y., Zhao, Y., Callcut, R. & Petzold, L. Integrating Physiological Time Series and Clinical Notes with Transformer for Early Prediction of Sepsis, DOI: 10.48550/arXiv.2203.14469 (2022). ArXiv:2203.14469 [cs].

57. Dalal, S., Ardabili, A. K. & Bonavia, A. S. Time-Series Deep Learning and Conformal Prediction for Improved Sepsis Diagnosis in Non-ICU Hospitalized Patients. *medRxiv* 2024.11.21.24317716, DOI: 10.1101/2024.11.21.24317716 (2024).

58. Ebadollahi, S. *et al.* Predicting Patient's Trajectory of Physiological Data using Temporal Trends in Similar Patients: A System for Near-Term Prognostics. *AMIA Annu. Symp. Proc.* **2010**, 192–196 (2010).

59. Rosenbaum, P. R. & Rubin, D. B. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **70**, 41–55, DOI: 10.2307/2335942 (1983). Publisher: [Oxford University Press, Biometrika Trust].

60. Shashikumar, S. P., Josef, C. S., Sharma, A. & Nemati, S. DeepAISE – An interpretable and recurrent neural survival model for early prediction of sepsis. *Artif. Intell. Medicine* **113**, 102036, DOI: 10.1016/j.artmed.2021.102036 (2021).

61. Moor, M. *et al.* Predicting sepsis using deep learning across international sites: a retrospective development and validation study. *EClinicalMedicine* **62**, 102124, DOI: 10.1016/j.eclinm.2023.102124 (2023).

Acknowledgements

ChatGPT and Claude were used for code debugging. ChatGPT and DeepL were employed for assistance with writing to improve grammar, vocabulary, and fluency.

Author contributions statement

Must include all authors, identified by initials, for example: A.A. conceived the experiment(s), A.A. and B.A. conducted the experiment(s), C.A. and D.A. analysed the results. All authors reviewed the manuscript.

Appendix

Supplemental Figures and Tables

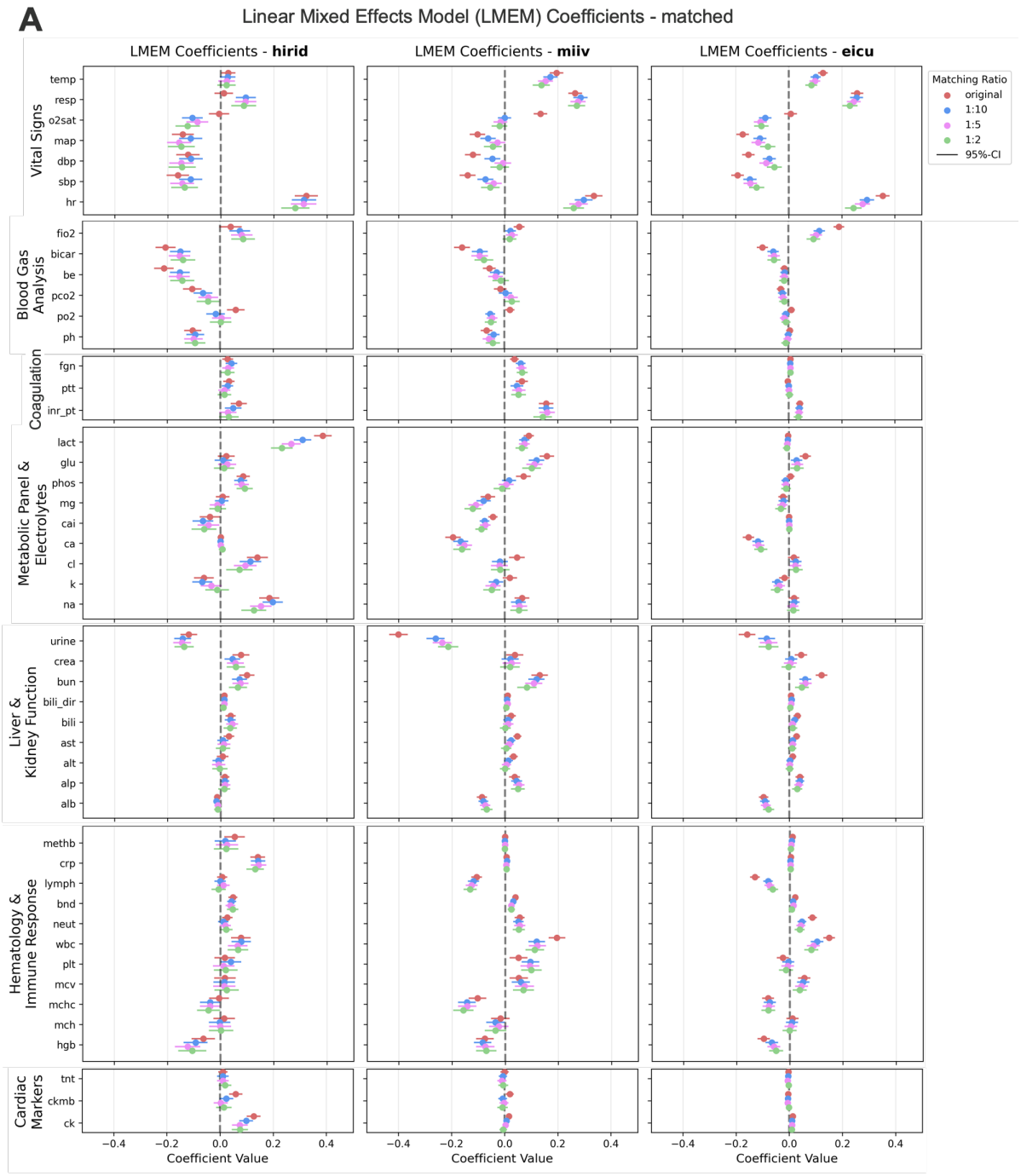


Figure 6. Absolute LMEM coefficients across datasets (original vs. matched). A) LMEM coefficients exhibit consistent patterns for dynamic features across all three datasets with higher magnitudes in original versus matched training cohorts, indicating attenuated discriminative capacity between sepsis and non-sepsis groups. Coefficients are plotted on the same scale for all grouped dynamic features and for all matched training cohort matching ratios. Positive coefficients indicate elevated feature values in the sepsis group relative to the non-sepsis group, whereas negative coefficients signify reduced values in sepsis cases. A coefficient of 1.0 denotes that the sepsis group exhibits, on average, a one-unit increase in the feature value compared to controls, after accounting for within-subject correlations via the mixed-effects model. All coefficients are presented with their respective 95% confidence intervals.

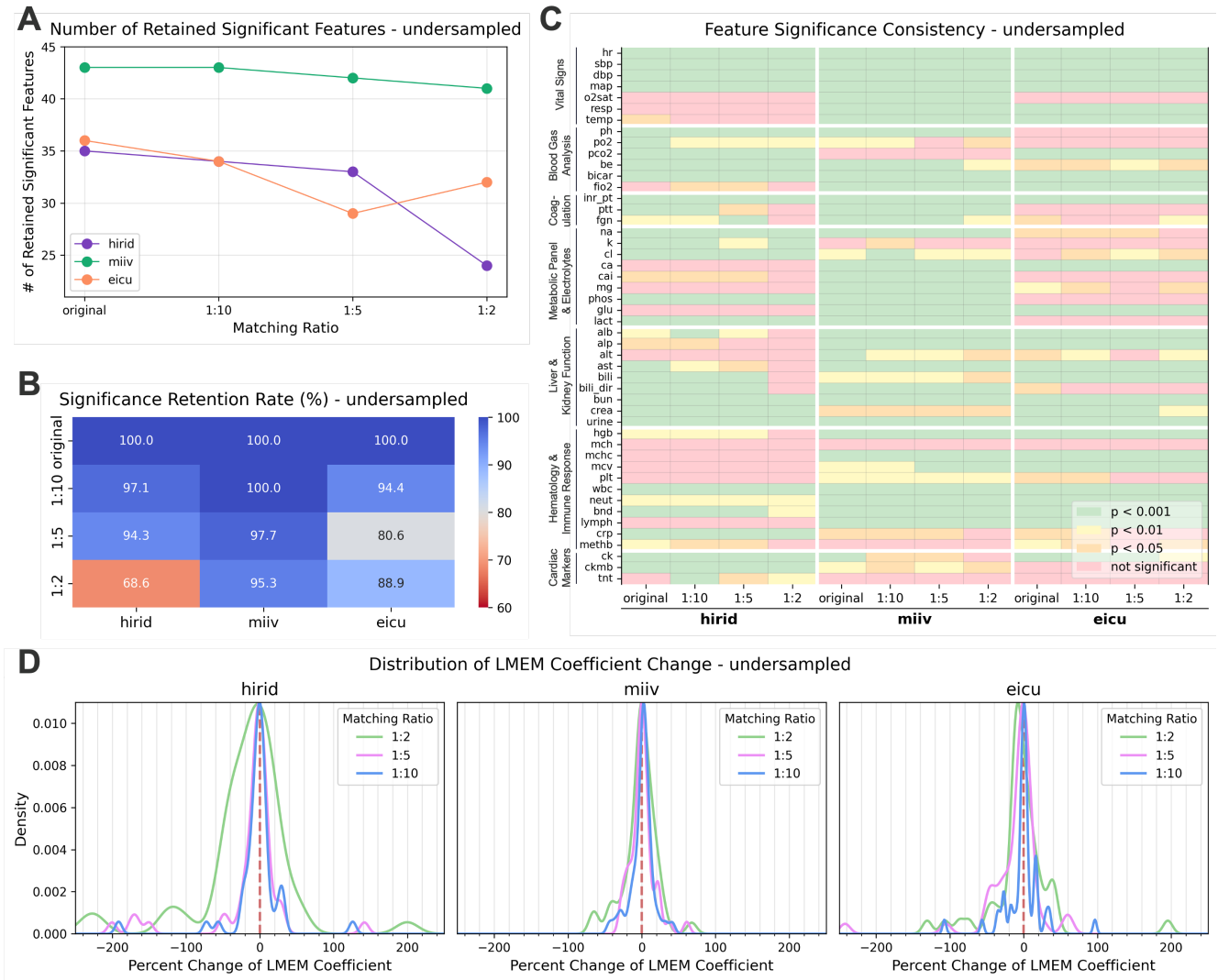


Figure 7. Evaluation of LMEM coefficient alterations and their statistical significance (original vs. undersampled). A) The absolute count of retained significant features in original versus undersampled training cohorts exhibits minimal reduction at 1:10 and 1:5 matching ratios. At 1:2 matching ratios, dataset-specific patterns emerge: MIMIC-IV and, to a slightly lesser extent, eICU preserve nearly all significant features, whereas HiRID demonstrates a substantial decrease. **B)** The relative significance retention rate similarly demonstrates dataset-specific patterns, with MIMIC-IV maintaining high preservation and HiRID exhibiting diminished retention at the 1:2 matching ratio. **C)** Analysis of significance levels reveals distinct heterogeneity in feature significance profiles across datasets. **D)** The percent change distribution in LMEM coefficients clusters around the original baseline with predominantly minor deviations, suggesting preserved discriminative capacity of collective dynamic features for sepsis versus non-sepsis cohort differentiation.

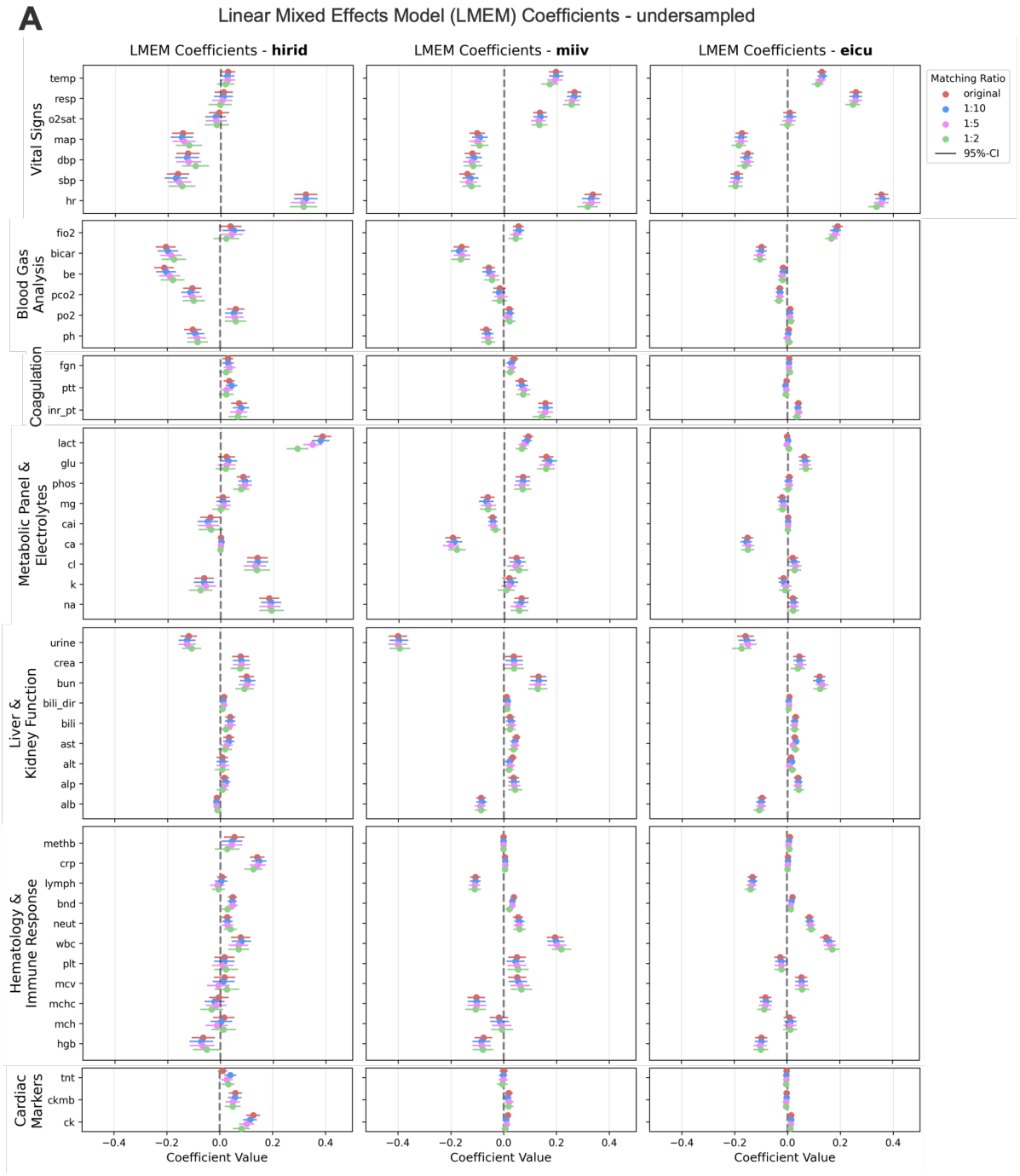


Figure 8. Absolute LMEM coefficients across datasets (original vs. undersampled). A) LMEM coefficients demonstrate negligible variation between original and undersampled training cohorts. Coefficients are plotted on the same scale for all grouped dynamic features and for all undersampled training cohort matching ratios. Positive coefficients indicate elevated feature values in the sepsis group relative to the non-sepsis group, whereas negative coefficients signify reduced values in sepsis cases. A coefficient of 1.0 denotes that the sepsis group exhibits, on average, a one-unit increase in the feature value compared to controls, after accounting for within-subject correlations via the mixed-effects model. All coefficients are presented with their respective 95% confidence intervals.

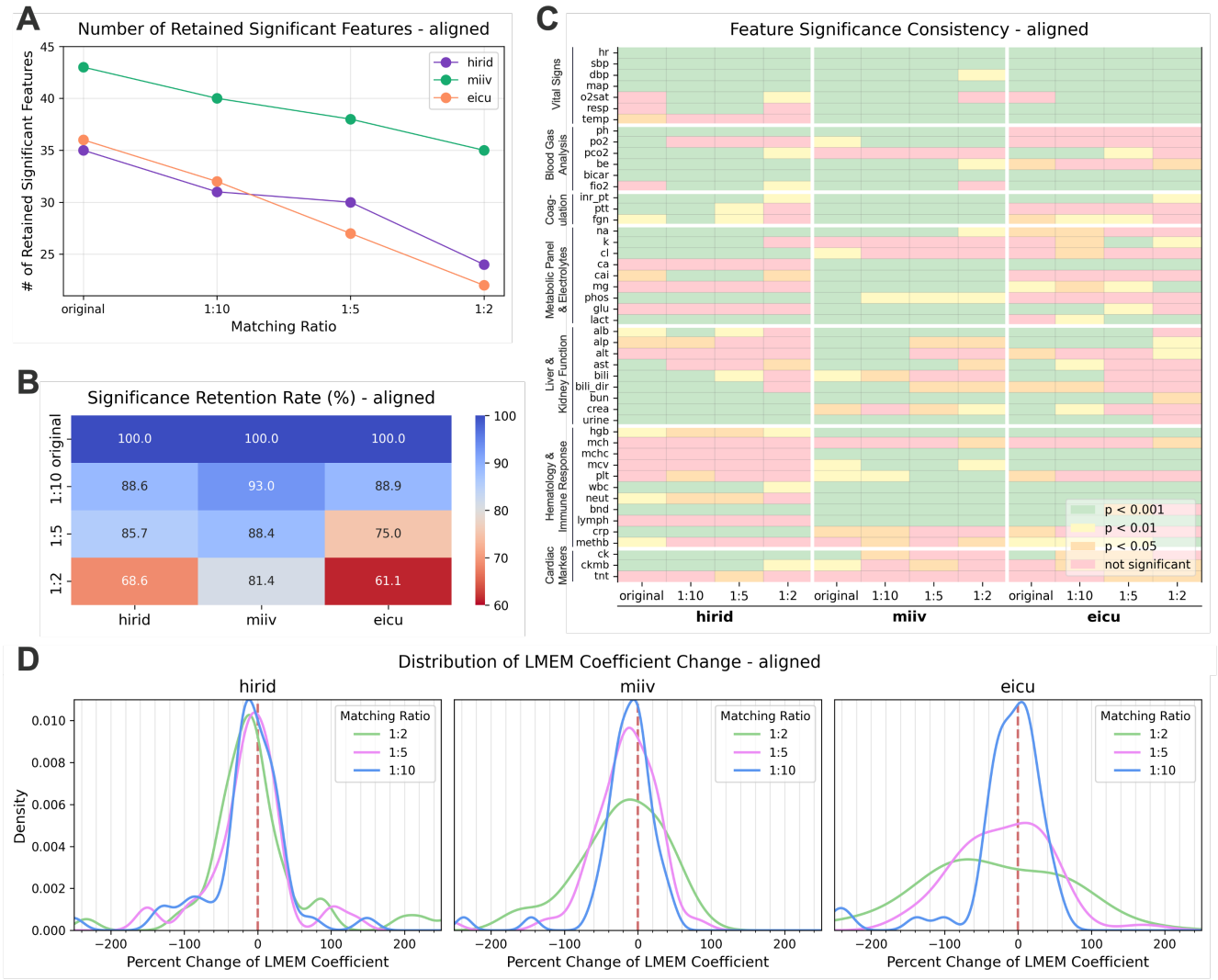


Figure 9. Evaluation of LMEM coefficient alterations and their statistical significance (original vs. aligned). A) The absolute number of retained significant features in aligned training cohorts compared to original training sets demonstrates an inverse relationship with increasing matching ratio. This relationship is comparable to results from matched training cohorts. **B)** The relative significance retention rate demonstrates attenuated decrements with increasing matching ratios in MIMIC-IV compared to HiRID and eICU. **C)** Analysis of significance levels reveals distinct heterogeneity in feature significance profiles across datasets. **D)** The distribution shift toward negative percent changes in LMEM coefficients indicates reduced discriminative capacity of the collective dynamic features for differentiating between sepsis and non-sepsis cohorts. Percent change values exhibited greater dispersion for eICU matching ratios of 1:5 and 1:2.

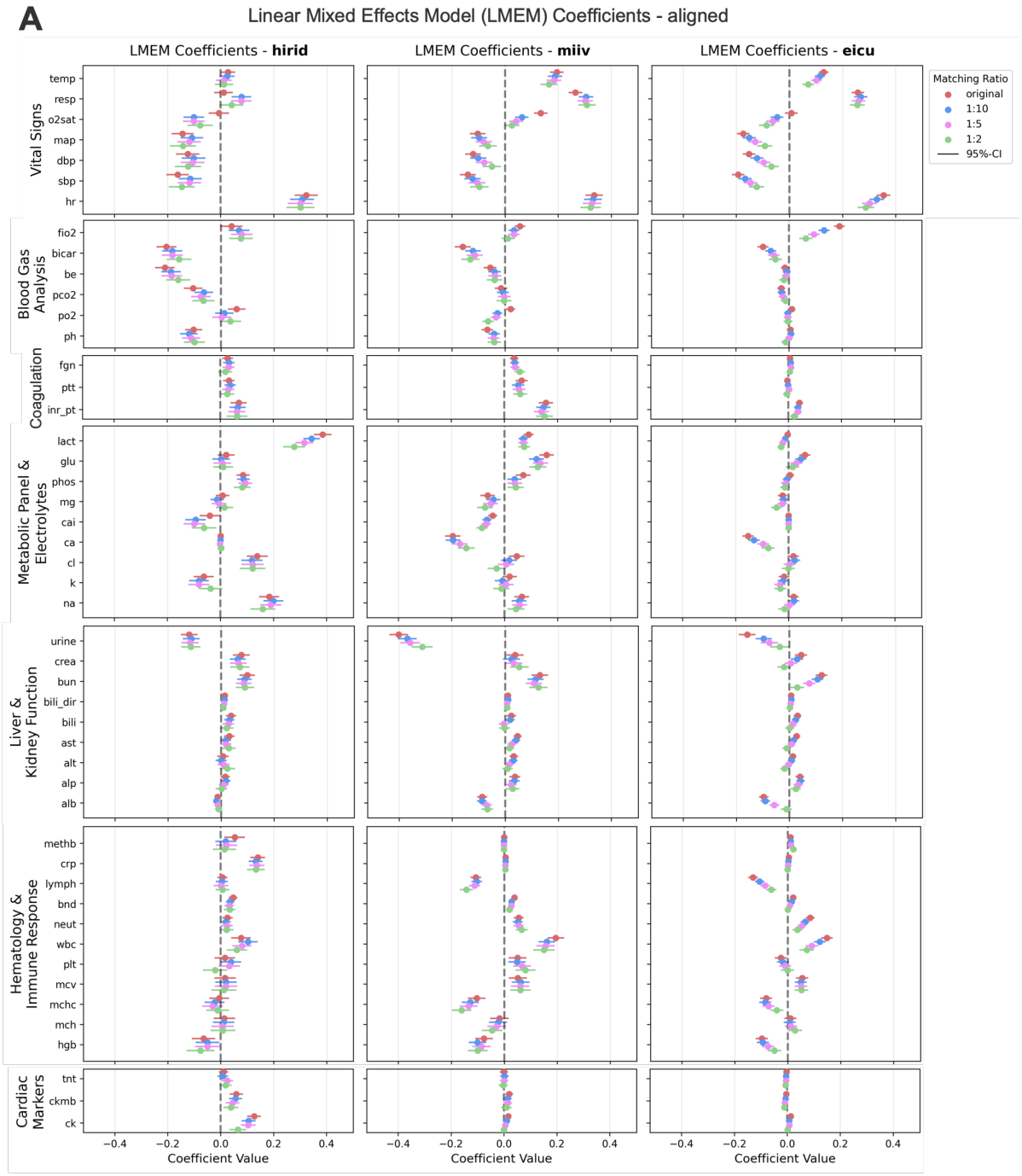


Figure 10. Absolute LMEM coefficients across datasets (original vs. aligned). A) LMEM coefficients parallel those derived from LMEM analysis of matched training cohorts. Coefficients are plotted on the same scale for all grouped dynamic features and for all aligned training cohort matching ratios. Positive coefficients indicate elevated feature values in the sepsis group relative to the non-sepsis group, whereas negative coefficients signify reduced values in sepsis cases. A coefficient of 1.0 denotes that the sepsis group exhibits, on average, a one-unit increase in the feature value compared to controls, after accounting for within-subject correlations via the mixed-effects model. All coefficients are presented with their respective 95% confidence intervals.

Category		Feature	ricu	*unit*	Reference Range		Thinkable Range*	
					lower	upper	lower	upper
Static Features		sex	sex	*M/F*				
		age	age	*years*				
		height	height	*cm*			135	225
		weight	weight	*kg*			40	250
Dynamic Features	Vital Signs	Heart Rate	hr	*bpm*	60	100	20	320
		Systolic Blood Pressure	sbp	*mmHg*	90	120	30	300
		Diastolic Blood Pressure	dbp	*mmHg*	60	80	10	200
		Mean Arterial Pressure (MAP)	map	*mmHg*	65	100	20	250
		Oxygen Saturation	o2sat	*%*	95	100	50	100
		Respiratory Rate	resp	*/min*	12	20	4	80
		Temperature	temp	*°C*	36.5	37.5	30	42
	Blood Gas Analysis	pH Level	ph	–	7.35	7.45	6.7	8
		Partial Pressure of Oxygen (PaO2)	po2	*mmHg*	75	100	40	600
		Partial Pressure of Carbon Dioxide (PaCO2)	pco2	*mmHg*	35	45	10	150
		Base Excess	be	*mmol/L*	-2	2	-25	25
		Bicarbonate	bicar	*mmol/L*	22	29	5	50
		Fraction of Inspired Oxygen (FiO2)	fio2	*%*	21	100	21	100
	Coagulation	International Normalized Ratio (INR)	inr_pt	–	0.8	1.2	0.5	20
		Partial Thromboplastin Time (PTT)	ptt	*sec*	25	35	10	250
		Fibrinogen	fgn	*mg/dL*	200	400	30	1100
	Metabolic Panel & Electrolytes	Sodium	na	*mmol/L*	135	145	90	170
		Potassium	k	*mmol/L*	3.5	5	1	9
		Chloride	cl	*mmol/L*	96	106	70	140
		Calcium	ca	*mg/dL*	8.5	10.5	4	20
		Ionized Calcium	cai	*mmol/L*	1.1	1.3	0.4	2.2
		Magnesium	mg	*mg/dL*	1.7	2.2	0.5	5
		Phosphate	phos	*mg/dL*	2.5	4.5	0.5	15
		Glucose	glu	*mg/dL*	70	140	25	1000
		Lactate	lact	*mmol/L*	0.5	2	0.1	20
	Liver & Kidney Function	Albumin	alb	*g/dL*	3.5	5	0.5	6
		Alkaline Phosphatase	alp	*U/L*	44	147	10	1200
		Alanine Aminotransferase (ALT)	alt	*U/L*	7	56	10	5000
		Aspartate Aminotransferase (AST)	ast	*U/L*	10	40	10	8000
		Total Bilirubin	bili	*mg/dL*	0.1	1.2	0.1	50
		Direct Bilirubin	bili_dir	*mg/dL*	0	0.3	0	30
		Blood Urea Nitrogen (BUN)	bun	*mg/dL*	7	20	1	180
		Creatinine	crea	*mg/dL*	0.6	1.3	0.1	20
	Hematology & Immune Response	Hemoglobin	hgb	*g/dL*	13.5	17.5	3	20
		Mean Corpuscular Hemoglobin (MCH)	mch	*pg*	27	33	15	45
		Mean Corpuscular Hemoglobin Concentration (MCHC)	mchc	*g/dL*	32	36	20	45
		Mean Corpuscular Volume (MCV)	mcv	*fL*	80	100	50	130
		Platelets	plt	$10^3/\mu L$	150	450	10	1500
		White Blood Cell Count (WBC)	wbc	$10^3/\mu L$	4	11	0.1	500
		Neutrophils	neut	*%*	55	70	0	100
		Band Neutrophils	bnd	*%*	0	6	0	50
		Lymphocytes	lymph	*%*	20	40	0	90
		C-Reactive Protein (CRP)	crp	*mg/L*	0	10	0	500
		Methemoglobin	methb	*%*	0	2	0	60
	Cardiac Markers	Creatine Kinase (CK)	ck	*U/L*	30	200	10	100000
		Creatine Kinase-MB (CK-MB)	ckmb	*ng/mL*	0	5	0	500
		Troponin T	tnt	*ng/mL*	0	14	0	1000

Table 3. Overview of features in harmonized datasets. Clinical concepts were extracted using `ricu`. Reference and thinkable ranges were defined after harmonization to facilitate data interpretation and cleaning. *Ranges were defined with generous boundaries and with respect to the circumstances of ICU stays. Outliers that were under or over the limits by more than 5% of the thinkable range were replaced with NaNs.

A — **Model Performance (vanilla)**

			AUROC										AUPRC									
			original	w/o demogr	undersampling			absolute onset matching			absolute onset matching + demographics criteria		original	w/o demogr	undersampling			absolute onset matching			absolute onset matching + demographics criteria		
					1:10	1:5	1:2	1:10	1:5	1:2	1:10	1:5	1:2		1:10	1:5	1:2	1:10	1:5	1:2	1:10	1:5	1:2
A1 1h data window, 6h prediction horizon	hirid	RF	0,727	0,727	0,988	0,976	0,978	0,461	0,345	0,357	0,405	0,346	0,319	0,061	0,058	0,950	0,899	0,881	0,020	0,025	0,037	0,018	0,018	0,028
		bRF	0,774	0,773	0,983	0,975	0,981	0,423	0,393	0,383	0,441	0,404	0,351	0,070	0,067	0,916	0,881	0,883	0,018	0,024	0,034	0,019	0,020	0,028
		LGBM	0,760	0,757	0,894	0,887	0,867	0,428	0,435	0,465	0,447	0,422	0,450	0,060	0,060	0,319	0,202	0,148	0,020	0,021	0,023	0,019	0,019	0,020
		XGB	0,720	0,732	0,915	0,917	0,894	0,422	0,439	0,481	0,442	0,430	0,446	0,062	0,065	0,416	0,338	0,188	0,018	0,023	0,026	0,019	0,020	0,020
	miiv	RF	0,711	0,705	0,989	0,990	0,970	0,372	0,320	0,308	0,389	0,370	0,331	0,029	0,029	0,939	0,948	0,864	0,010	0,010	0,019	0,010	0,015	0,021
		bRF	0,775	0,776	0,986	0,988	0,973	0,460	0,415	0,375	0,474	0,439	0,377	0,035	0,036	0,909	0,919	0,826	0,011	0,012	0,021	0,011	0,015	0,017
		LGBM	0,773	0,761	0,871	0,863	0,835	0,467	0,475	0,487	0,451	0,472	0,469	0,034	0,031	0,148	0,098	0,073	0,011	0,011	0,012	0,010	0,012	0,012
		XGB	0,755	0,743	0,908	0,901	0,867	0,462	0,458	0,455	0,460	0,442	0,472	0,034	0,035	0,269	0,170	0,098	0,011	0,011	0,011	0,011	0,012	0,012
	eicu	RF	0,619	0,629	0,975	0,966	0,969	0,367	0,328	0,332	0,397	0,338	0,337	0,017	0,018	0,906	0,885	0,891	0,010	0,011	0,016	0,010	0,012	0,024
		bRF	0,692	0,688	0,969	0,967	0,970	0,471	0,435	0,383	0,477	0,426	0,383	0,025	0,024	0,878	0,851	0,865	0,012	0,015	0,018	0,011	0,013	0,026
		LGBM	0,697	0,703	0,794	0,784	0,779	0,494	0,493	0,483	0,494	0,475	0,490	0,025	0,026	0,085	0,059	0,047	0,011	0,011	0,011	0,011	0,010	0,012
		XGB	0,675	0,677	0,846	0,835	0,816	0,507	0,488	0,483	0,496	0,478	0,485	0,023	0,024	0,119	0,106	0,057	0,012	0,011	0,011	0,011	0,011	0,011
A2 6h data window, 6h prediction horizon	hirid	RF	0,735	0,733	0,989	0,973	0,978	0,483	0,367	0,378	0,425	0,405	0,374	0,069	0,069	0,945	0,895	0,871	0,023	0,024	0,040	0,019	0,021	0,032
		bRF	0,784	0,781	0,985	0,974	0,979	0,468	0,409	0,371	0,464	0,431	0,398	0,078	0,079	0,921	0,861	0,812	0,022	0,024	0,036	0,021	0,023	0,036
		LGBM	0,785	0,773	0,914	0,906	0,889	0,452	0,429	0,442	0,478	0,475	0,460	0,086	0,081	0,399	0,287	0,204	0,021	0,025	0,026	0,023	0,023	0,022
		XGB	0,754	0,760	0,934	0,931	0,932	0,423	0,409	0,450	0,478	0,462	0,435	0,086	0,082	0,498	0,415	0,322	0,019	0,020	0,027	0,022	0,021	0,023
	miiv	RF	0,763	0,763	0,991	0,990	0,972	0,403	0,346	0,345	0,428	0,400	0,350	0,043	0,044	0,942	0,944	0,846	0,010	0,010	0,020	0,010	0,015	0,021
		bRF	0,812	0,810	0,987	0,990	0,972	0,493	0,460	0,400	0,487	0,454	0,422	0,049	0,050	0,888	0,852	0,689	0,012	0,012	0,021	0,012	0,015	0,019
		LGBM	0,817	0,816	0,901	0,910	0,883	0,474	0,477	0,488	0,448	0,434	0,475	0,048	0,048	0,190	0,143	0,098	0,011	0,011	0,011	0,010	0,011	0,012
		XGB	0,802	0,803	0,935	0,926	0,901	0,467	0,458	0,470	0,455	0,450	0,460	0,052	0,055	0,318	0,232	0,148	0,011	0,010	0,011	0,011	0,012	0,012
	eicu	RF	0,657	0,655	0,975	0,970	0,972	0,397	0,367	0,368	0,426	0,367	0,362	0,022	0,023	0,905	0,885	0,883	0,010	0,012	0,017	0,010	0,014	0,024
		bRF	0,730	0,730	0,968	0,970	0,971	0,527	0,482	0,424	0,520	0,469	0,435	0,033	0,033	0,835	0,781	0,734	0,014	0,016	0,020	0,013	0,016	0,027
		LGBM	0,740	0,733	0,820	0,820	0,812	0,512	0,507	0,481	0,498	0,485	0,498	0,037	0,035	0,062	0,072	0,067	0,012	0,012	0,011	0,011	0,011	0,012
		XGB	0,721	0,716	0,876	0,866	0,860	0,510	0,499	0,499	0,493	0,472	0,491	0,034	0,034	0,174	0,123	0,094	0,012	0,012	0,012	0,011	0,011	0,012
A3 6h data window, 10h prediction horizon	hirid	RF	0,716	0,706	0,986	0,975	0,976	0,479	0,339	0,361	0,419	0,398	0,385	0,064	0,065	0,941	0,895	0,872	0,022	0,020	0,044	0,019	0,021	0,031
		bRF	0,762	0,759	0,982	0,971	0,974	0,473	0,425	0,391	0,474	0,433	0,420	0,075	0,071	0,911	0,828	0,722	0,023	0,022	0,037	0,021	0,026	0,033
		LGBM	0,773	0,773	0,926	0,913	0,901	0,446	0,440	0,449	0,478	0,446	0,482	0,073	0,083	0,383	0,271	0,206	0,019	0,024	0,021	0,022	0,020	0,023
		XGB	0,723	0,738	0,952	0,934	0,928	0,440	0,436	0,483	0,448	0,464	0,484	0,085	0,085	0,586	0,402	0,312	0,019	0,023	0,031	0,020	0,022	0,027
	miiv	RF	0,732	0,733	0,990	0,988	0,966	0,393	0,340	0,336	0,415	0,405	0,340	0,037	0,034	0,940	0,937	0,839	0,009	0,011	0,020	0,009	0,013	0,018
		bRF	0,783	0,781	0,985	0,988	0,961	0,509	0,446	0,413	0,489	0,452	0,405	0,042	0,042	0,856	0,776	0,519	0,012	0,011	0,022	0,012	0,014	0,016
		LGBM	0,785	0,773	0,903	0,907	0,881	0,479	0,499	0,497	0,449	0,458	0,486	0,037	0,036	0,172	0,130	0,092	0,010	0,012	0,010	0,009	0,010	0,011
		XGB	0,759	0,747	0,938	0,944	0,900	0,472	0,468	0,485	0,479	0,460	0,480	0,039	0,040	0,338	0,257	0,112	0,010	0,011	0,011	0,010	0,010	0,011
	eicu	RF	0,646	0,645	0,971	0,970	0,970	0,388	0,366	0,350	0,427	0,354	0,357	0,021	0,021	0,906	0,887	0,882	0,010	0,012	0,016	0,010	0,012	0,023
		bRF	0,709	0,709	0,964	0,968	0,967	0,519	0,492	0,433	0,531	0,476	0,442	0,031	0,030	0,813	0,722	0,606	0,014	0,017	0,021	0,013	0,017	0,027
		LGBM	0,702	0,707	0,813	0,817	0,815	0,520	0,511	0,488	0,519	0,495	0,491	0,029	0,030	0,083	0,072	0,061	0,012	0,011	0,011	0,012	0,011	0,011
		XGB	0,694	0,700	0,884	0,875	0,857	0,519	0,507	0,488	0,502	0,504	0,490	0,030	0,031	0,169	0,132	0,093	0,011	0,013	0,011	0,011	0,012	0,011

AUROC: 0,31 – 0,99 AUPRC: 0,00 – 0,95

Figure 11. Performance of machine learning models. A) Comparative analysis of non-optimized (vanilla) machine learning models trained on specified training datasets (original, demographics-excluded, as well as undersampled, aligned, and matched at 1:10/1:5/1:2 ratios each) and evaluated on the original test cohort. Left panels display AUROC metrics while right panels present AUPRC metrics. Note that AUPRC values cannot be directly compared due to varying positive-class prevalence and therefore baseline performance across training cohorts (Table 5, Figure 12). This experimental framework was implemented across three distinct temporal configurations: **(A1)** 1-hour data window with 0-hour prediction window, **(A2)** 6-hour data window with 0-hour prediction window, and **(A3)** 6-hour data window with 4-hour prediction window. Models trained on undersampled cohorts demonstrate substantial performance metric improvements, while those trained on absolute-onset matched cohorts—with or without demographic criteria—exhibit performance metrics below baseline levels.

			Models Analyzed (Σ 108)		Metric Change	
			Count	**Percentage**	**Mean**	**Standard Deviation**
Differences in Model Performance (Matched vs. Aligned)	AUROC	↑/=	58	54%	0,019	0,017
		↓	50	46%	-0,017	0,014
	nAUPRC	↑/=	54	50%	0,025	0,030
		↓	54	50%	-0,040	0,027
	AUPRC	↑/=	63	58%	0,002	0,002
		↓	45	42%	-0,003	0,003

Table 4. Differences in model performance metrics between matched and aligned cohorts. Arrows indicate performance improvement (↑), no change (=), or degradation (↓) of models trained on matched training sets (across all matching ratios) when compared to models trained on aligned training sets. A comprehensive evaluation was conducted across 108 distinct models, comprising a factorial design of 3 datasets, 4 model algorithms, 3 data-prediction-window configurations, and 3 matching ratios. The comparative analysis of matched versus aligned sampling methods reveals that the incorporation of additional demographic matching criteria produces minimal impact on model performance. The metrics AUROC, nAUPRC, and AUPRC demonstrate comparable magnitudes of change, with approximately half of the models showing slight improvements and half showing slight decrements in performance across each metric.

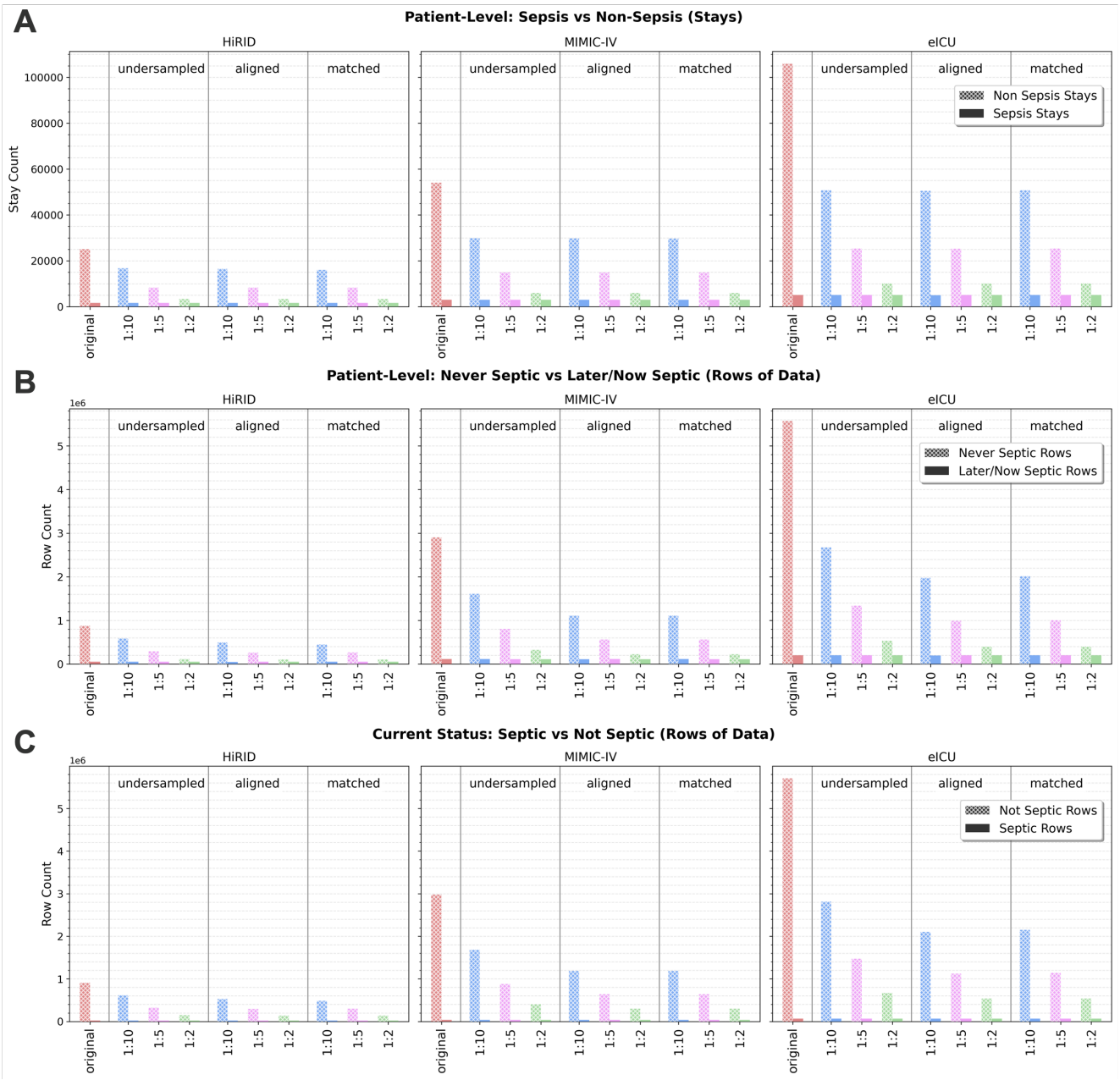


Figure 12. Overview of training set sizes and sepsis case and label prevalences. Implementation of three case-control matching ratios (1:10, 1:5, and 1:2) yields progressively reduced training cohorts with inversely increasing sepsis prevalence. Numerical data for these distributions are presented in Table 5. **A)** Number of stays in the sepsis and non-sepsis groups across different datasets (HiRID, MIMIC-IV, eICU), matching methods (undersampled, aligned, matched) and matching ratios (1:10, 1:5, 1:2). **B)** Combined number of rows (i.e. hours of data) of sepsis (later or now septic) and non-sepsis (never septic) stays. **B)** Number of rows (i.e. hours of data) with positive sepsis label (currently septic) and negative sepsis label (currently not septic).

Dataset	Matching Method	Variant	Set	Total Stays	Non-Sepsis Cases	Sepsis Cases (Prevalence)
HiRID	original		train	26,728	25,040 (93.7%)	1,688 (6.3%)
	original		test	2,970	2,800 (94.3%)	170 (5.7%)
	undersampled	10	train	18,394	16,726 (90.9%)	1,668 (9.1%)
	undersampled	5	train	10,023	8,354 (83.3%)	1,679 (16.7%)
	undersampled	2	train	5,016	3,343 (66.6%)	1,673 (33.4%)
	aligned	10	train	18,394	16,726 (90.9%)	1,668 (9.1%)
	aligned	5	train	10,011	8,350 (83.2%)	1,668 (16.8%)
	aligned	2	train	5,016	3,343 (66.6%)	1,673 (33.4%)
	matched	10	train	17,681	16,013 (90.6%)	1,668 (9.4%)
	matched	5	train	10,028	8,354 (83.3%)	1,674 (16.7%)
	matched	2	train	5,069	3,395 (67.0%)	1,674 (33.0%)
MIMIC-IV	original		train	57,082	54,058 (94.7%)	3,024 (5.3%)
	original		test	6,343	6,046 (95.3%)	297 (4.7%)
	undersampled	10	train	32,877	29,874 (90.9%)	3,003 (9.1%)
	undersampled	5	train	17,933	14,933 (83.3%)	3,000 (16.7%)
	undersampled	2	train	8,966	5,969 (66.6%)	2,997 (33.4%)
	aligned	10	train	32,751	29,765 (90.9%)	2,986 (9.1%)
	aligned	5	train	17,933	14,934 (83.3%)	2,999 (16.7%)
	aligned	2	train	8,965	5,969 (66.6%)	2,996 (33.4%)
	matched	10	train	32,734	29,737 (90.8%)	2,998 (9.2%)
	matched	5	train	17,933	14,946 (83.3%)	2,987 (16.7%)
	matched	2	train	8,966	5,969 (66.6%)	2,997 (33.4%)
eICU	original		train	111,071	105,988 (95.4%)	5,083 (4.6%)
	original		test	12,342	11,787 (95.5%)	555 (4.5%)
	undersampled	10	train	55,816	50,739 (90.9%)	5,077 (9.1%)
	undersampled	5	train	30,445	25,363 (83.3%)	5,082 (16.7%)
	undersampled	2	train	15,222	10,143 (66.6%)	5,079 (33.4%)
	aligned	10	train	55,539	50,522 (91.0%)	5,017 (9.0%)
	aligned	5	train	30,522	25,341 (83.0%)	5,181 (17.0%)
	aligned	2	train	15,222	10,141 (66.6%)	5,081 (33.4%)
	matched	10	train	55,816	50,734 (90.9%)	5,082 (9.1%)
	matched	5	train	30,445	25,363 (83.3%)	5,082 (16.7%)
	matched	2	train	15,222	10,155 (66.7%)	5,067 (33.3%)

Dataset	Matching Method	Variant	Set	Total Rows	Group		Label	
					Never-Septic	Later-/Now-Septic	Not Septic	Septic
HiRID	original		train	930,608	877,612 (94.3%)	52,996 (5.7%)	908,712 (97.6%)	21,896 (2.4%)
	original		test	102,473	96,175 (93.9%)	6,298 (6.1%)	100,267 (97.8%)	2,206 (2.2%)
	undersampled	10	train	504,536	451,320 (89.5%)	53,220 (10.5%)	482,883 (95.7%)	21,754 (4.3%)
	undersampled	5	train	319,754	266,046 (83.2%)	53,708 (16.8%)	297,983 (93.2%)	21,771 (6.8%)
	undersampled	2	train	159,190	105,281 (66.1%)	53,909 (33.9%)	137,402 (86.3%)	21,788 (13.7%)
	aligned	10	train	519,289	466,172 (89.8%)	53,126 (10.2%)	497,492 (95.8%)	21,798 (4.2%)
	aligned	5	train	328,286	266,859 (81.3%)	61,432 (18.7%)	306,455 (93.3%)	21,837 (6.7%)
	aligned	2	train	159,076	105,229 (66.1%)	53,848 (33.9%)	137,284 (86.3%)	21,802 (13.7%)
	matched	10	train	504,519	451,264 (89.4%)	53,255 (10.6%)	482,883 (95.7%)	21,636 (4.3%)
	matched	5	train	319,754	266,046 (83.2%)	53,708 (16.8%)	298,047 (93.2%)	21,714 (6.8%)
	matched	2	train	159,404	105,563 (66.2%)	53,891 (33.8%)	137,598 (86.3%)	21,836 (13.7%)
MIMIC-IV	original		train	3,017,646	2,903,317 (96.2%)	114,329 (3.8%)	2,978,396 (98.7%)	39,250 (1.3%)
	original		test	341,958	330,615 (96.7%)	11,343 (3.3%)	338,099 (98.9%)	3,859 (1.1%)
	undersampled	10	train	1,722,288	1,608,255 (93.4%)	114,037 (6.6%)	1,683,395 (97.7%)	38,896 (2.3%)
	undersampled	5	train	978,358	865,528 (88.5%)	112,832 (11.5%)	939,470 (96.0%)	38,893 (4.0%)
	undersampled	2	train	489,048	376,253 (76.9%)	112,795 (23.1%)	450,127 (92.0%)	38,942 (8.0%)
	aligned	10	train	1,272,709	1,152,570 (90.6%)	119,530 (9.4%)	1,234,222 (96.9%)	38,528 (3.1%)
	aligned	5	train	697,878	584,053 (83.6%)	114,440 (16.4%)	659,128 (94.3%)	39,570 (5.7%)
	aligned	2	train	337,942	225,240 (66.6%)	113,034 (33.4%)	299,211 (88.5%)	38,795 (11.5%)
	matched	10	train	1,223,350	1,109,025 (90.7%)	114,324 (9.3%)	1,184,435 (96.8%)	38,915 (3.2%)
	matched	5	train	678,271	565,436 (83.4%)	112,889 (16.6%)	639,377 (94.3%)	38,910 (5.7%)
	matched	2	train	336,483	224,231 (66.6%)	112,252 (33.4%)	297,583 (88.4%)	38,900 (11.6%)
eICU	original		train	5,772,659	5,571,761 (96.5%)	200,898 (3.5%)	5,707,176 (98.9%)	65,483 (1.1%)
	original		test	644,384	621,458 (96.4%)	22,926 (3.6%)	637,231 (98.9%)	7,153 (1.1%)
	undersampled	10	train	2,575,364	2,370,363 (92.0%)	205,001 (8.0%)	2,512,803 (97.6%)	62,561 (2.4%)
	undersampled	5	train	1,530,332	1,330,332 (86.9%)	200,001 (13.1%)	1,467,619 (95.9%)	62,713 (4.1%)
	undersampled	2	train	737,419	535,312 (72.6%)	202,204 (27.4%)	671,967 (91.1%)	65,432 (8.9%)
	aligned	10	train	1,609,570	1,375,572 (85.5%)	198,653 (14.5%)	1,514,923 (94.8%)	84,647 (5.2%)
	aligned	5	train	920,273	721,472 (78.4%)	200,503 (21.6%)	838,823 (91.0%)	83,064 (9.0%)
	aligned	2	train	602,885	400,605 (66.5%)	202,349 (33.5%)	537,348 (90.0%)	65,484 (10.0%)
	matched	10	train	2,214,452	2,013,255 (90.9%)	201,197 (9.1%)	2,148,990 (97.0%)	65,462 (3.0%)
	matched	5	train	1,207,335	1,004,836 (83.2%)	202,499 (16.8%)	1,141,852 (94.6%)	65,483 (5.4%)
	matched	2	train	603,836	402,276 (66.6%)	201,560 (33.4%)	538,555 (89.2%)	65,281 (10.8%)

Table 5. Summary of training and testing set characteristics across different matching methods and ratios. Machine learning analyses were conducted using the original dataset versions for both model training and evaluation, whereas the undersampled, aligned and matched dataset variants were exclusively utilized for model training. All models, regardless of training dataset, were subsequently evaluated on the original test sets to maintain external validity and enable assessment under clinically representative conditions. The total number of sepsis cases can vary slightly because of randomized splitting into training and testing sets for all variants. Visual representations of this data are depicted in Figure 12.

Model	Hyperparameter	Range	Best Parameters		
			HiRID	MIMIC-IV	eICU
RF	n_estimators	50-500	486	477	467
	max_depth	3-30	19	16	13
	min_samples_split	2-10	3	7	2
	min_samples_leaf	1-4	2	3	2
bRF	n_estimators	50-500	409	415	400
	max_depth	3-30	28	29	29
	min_samples_split	2-10	9	9	9
LGBM	n_estimators	50-500	81	76	489
	learning_rate	0.01-0.3	0.054	0.055	0.012
	max_depth	3-30	6	7	15
	num_leaves	20-100	92	89	62
XGB	n_estimators	50-500	188	179	478
	learning_rate	0.01-0.3	0.053	0.092	0.028
	max_depth	3-7	6	4	5
	min_child_weight	1-5	4	3	2

Table 6. Bayesian optimization hyperparameter spaces and best parameters. Hyperparameter ranges were defined for each machine learning model, with optimal parameters determined independently for each model-dataset pairing.

Code Availability

Scripts used to generate the results presented in this study are publicly available in a GitLab repository.
The following Python scripts are included:

1. **Data Preparation**

   - `preprocessing.py`
     This script utilizes harmonized datasets generated via the YAIB workflow (https://github.com/rvandewater/YAIB). It provides a flexible pipeline for specifying datasets (HiRID, MIIV, eICU), matching methods (original, undersampled, aligned, matched), matching ratios (1:2, 1:5, 1:10), and tolerances for strict demographics-based matching.

   - `windowing.py`
     This script processes the preprocessed training and testing sets (output from `preprocessing.py`) to generate windowed datasets. Users can select preprocessed datasets by specifying the dataset (HiRID, MIIV, eICU), matching method (original, undersampled, aligned, matched), and matching ratio (1:2, 1:5, 1:10). Data and prediction window sizes are configurable (default combinations: 1/0, 6/0, 6/4).

2. **Linear Mixed Effects Models**

   - `lmem.py`
     This script analyzes unwindowed preprocessed training sets (output from `preprocessing.py`) and trains univariate linear mixed effects models for each dynamic feature. Users can specify datasets (HiRID, MIIV, eICU) and matching methods (undersampled, aligned, matched), while original training sets are always included as a comparison. Output is provided as a text file report.

   - `lmem_visualizations.py`
     This script converts the text file report (output from `lmem.py`) into a pandas DataFrame and generates visualizations.

3. **Machine Learning Models**

   - `ml_vanilla.py`
     This script operates on preprocessed unwindowed (1-hour data window, 0-hour prediction window) or windowed training and testing sets (output from `preprocessing.py` and `windowing.py`). It offers a flexible framework for specifying datasets (HiRID, MIIV, eICU), machine learning algorithms (RF, bRF, LGBM, XGB), window configurations (default data-prediction-window combinations: 1/0, 6/0, 6/4), matching methods (original, undersampled, aligned, matched), and matching ratios (1:2, 1:5, 1:10). The script trains models and evaluates their performance.

   - `ml_bayesianopt.py`
     This script utilizes preprocessed unwindowed (1-hour data window, 0-hour prediction window) training and testing sets (output from `preprocessing.py`). Users can specify datasets (HiRID, MIIV, eICU) and machine learning algorithms (RF, bRF, LGBM, XGB). The script applies Bayesian optimization to machine learning models, identifies optimal hyperparameters, and evaluates the resulting model performance.