Flexible Analysis of Electronic Medical Record Data with Composite Mixture Models

Michael B. Mayhew[a,*], Brenden K. Petersen[a], Ana Paula Sales[a], John D. Greene[b], Vincent X. Liu[b], Todd S. Wasson[a,1]

[a] *Computational Engineering Division*
Lawrence Livermore National Laboratory
Livermore, California USA 94550
[b] *Division of Research*
Kaiser Permanente Northern California
Oakland, California USA 94611

Abstract

The widespread adoption of electronic medical records (EMRs) in healthcare has provided vast new amounts of data for statistical machine learning researchers in their efforts to model and predict patient health status, potentially enabling novel advances in treatment. However, there are significant barriers that must be overcome to extract these insights from EMR data. First, EMR datasets consist of both static and dynamic observations of discrete and continuous-valued variables, many of which may be missing, precluding the application of standard multivariate analysis techniques. Second, clinical populations observed via EMRs and relevant to the study and management of debilitating conditions like sepsis are often heterogeneous; properly accounting for this heterogeneity is critical. Here, we describe a joint probabilistic framework called a composite mixture model that can simultaneously accommodate the wide variety of observations frequently observed in EMR datasets, stratify heterogeneous clinical populations into relevant subgroups, and handle missing observations. We demonstrate the efficacy of our approach by applying our framework to a large-scale sepsis cohort, identifying physiological trends and distinct subgroups of the dataset associated with elevated risk of mortality during hospitalization.

Keywords: electronic health records, mixture modeling, risk stratification, sepsis, composite mixture model, cluster analysis

[*] Principal corresponding author

Email addresses: `mayhew5@llnl.gov` (Michael B. Mayhew), `petersen33@llnl.gov` (Brenden K. Petersen), `apsales@llnl.gov` (Ana Paula Sales), `john.d.greene@kp.org` (John D. Greene), `vincent.x.liu@kp.org` (Vincent X. Liu), `twasson@gmail.com` (Todd S. Wasson)

[1] Author since relocated to LinkedIn, Inc.

[2] **Abbreviations**: EMR - electronic medical record, CMM - composite mixture model, KPNC - Kaiser Permanente Northern California, BIC - Bayesian information criterion, AIC - Akaike information criterion, PAM - partitioning around medoids, MICE - multivariate imputation using chained equations

1. Introduction

Electronic medical records (EMRs) have become increasingly ubiquitous in healthcare, and the utility of these complex datasets for clinical decision support is the subject of much current research ([5] and references therein). EMR data comprise multivariate observations of variables with discrete or continuous values. These variables can be static (observed only a single time during a hospitalization; e.g. gender) or dynamic (e.g. vital signs). Deriving actionable insights from EMR data requires appropriate models for these multi-typed observations [19].

Besides complexity in the types of information contained in the EMR, heterogeneity inherent in the clinical population under study adds to the challenge of modeling these data. This physiological heterogeneity is a hallmark of debilitating conditions like cancer [11] and, in particular, sepsis [16], an increasingly prevalent clinical condition characterized by a dysregulated immune response to infection leading to organ dysfunction and death. Accounting for this heterogeneity can have considerable therapeutic importance. In the case of breast cancer, for example, stratification of patient tumors into molecular subtypes significantly increased precision of treatment and improved survival (reviewed in [24]). Indeed, a central goal of this and other similar efforts is to identify physiologically distinct subgroups of a clinical population more at risk for adverse health outcomes that can be targeted with interventions tailored to the subgroup's characteristics [15, 28, 20, 17, 29].

One approach to modeling EMR data focuses on the important task of prediction in order to prevent or mitigate an impending adverse outcome. Traditional predictive modeling approaches, such as logistic regression and random forests [3], can provide useful insights for clinical decision support particularly in identifying the constellation of patient features that indicate unfavorable health outcomes. For example, multi-task predictive models have proven successful in stratifying patients according to risk of developing hospital-acquired infections [28, 29]. Particularly in sepsis, these approaches have demonstrated remarkable performance in early identification of patients at risk of entering septic shock [12]. In essence, these models characterize the effects of a fixed set of predictor variables or features on an outcome of interest (*supervised* learning; a *conditional* model), without directly modeling the features themselves. These approaches also require complete case data (i.e. observations without missing entries). Without first performing imputation, certain samples or even entire features could be removed from analysis. More importantly, many of these approaches assume that the population under study can be modeled as a single, homogeneous group, ignoring the possibility that the population represents a mixture of distinct subgroups potentially amenable to different treatment regimes.

Unsupervised approaches, on the other hand, learn structure or dependencies by modeling the whole data observation (outcome and features) together,

capturing dependencies among the different dimensions of the full observation vector. Such approaches have proven successful recently in directly modeling complex EMR observations or embedding them in lower-dimensional feature spaces that can then be used for prediction or risk stratification [15, 20, 1, 17]. Multivariate statistical mixture models are another class of unsupervised learning approach that can formally represent heterogeneity in the population, treating both outcomes and features as probabilistic quantities to be modeled jointly (*joint* model; [23]). However, such mixture models are difficult to apply in their conventional form when elements of a multivariate observation are of different types (e.g. a categorical variable like gender and a continuous variable like median diastolic blood pressure) as is common with EMR data.

In this study, we describe a joint probabilistic framework called a composite mixture model (CMM; [22, 26]), a technique heretofore never applied to EMR data. The CMM accommodates the wide variety of data types common in EMR datasets while accounting for heterogeneity in the clinical population. We adapt our model to analyze a large EMR dataset composed of more than 53,000 emergency department hospitalization episodes from Kaiser Permanente Northern California (KPNC) in which patients were suspected to have infection and, in a subset of cases, met the criteria for sepsis [25, 6]. We demonstrate the efficacy of our approach by stratifying patient episodes into clusters with different levels of risk for mortality at the end of hospitalization, by visualizing physiological trends associated with higher rates of mortality, and by benchmarking the performance of our framework on common EMR analysis tasks such as missing data imputation.

2. Methods

2.1. KPNC Sepsis Cohort Description

Kaiser Permanente Northern California is a highly integrated healthcare delivery system with 21 medical centers caring for an overall population of 4 million members. The full KPNC dataset consists of 244,248 inpatient hospitalization episodes with a suspected or confirmed infection and sepsis diagnoses, drawn from KPNC medical centers between 2009 and 2013 [25]. This large-scale dataset consists of admission data as well as laboratory and vital sign values collected over the course of a patient's hospitalization. For this analysis, we used a subset of the full dataset consisting of 53,659 episodes, with an overall mortality rate of $\sim$6%. Our criteria for creating this analysis cohort were the following: 1) hospitalization admission occurred via the emergency department; 2) the length of hospitalization was at least twelve hours; and 3) all vital signs were taken three or more times during the first three hours of hospitalization.

Our cohort consisted of both static and dynamic features. The static features that made up our dataset include: 1) age; 2) sex; 3) treatment facility code; 4) KPNC membership status; 5) indicator of whether patient was transported in from another site; 6) LAPS2, a KPNC single measure of *acute* disease burden at the time of hospital admission; 7) COPS2, a monthly KPNC aggregate measure

of *chronic* disease burden; and 8) an indicator of patient mortality status at the end of the episode.

To take into account changes in patient physiology during hospitalization, we computed the maximum, minimum, median, and standard deviation of patient vital signs over different post-admission periods (3, 6, or 12 hours). Pairing the summary statistics with their corresponding static admission features resulted in three analysis cohorts, one for each post-admission period. The vital signs included in this analysis were 1) diastolic blood pressure, 2) systolic blood pressure, 3) heart rate, 4) respiratory rate, 5) temperature, and 6) pulse pressure.

2.2. Composite Mixture Model Definition

Here we describe the composite mixture model, a flexible joint probability model for multi-typed, multivariate data. The two central ideas behind the CMM are that: 1) the population is heterogeneous (composed of subgroups or clusters) and 2) we can specify the full joint distribution of a multi-typed observation vector, x, by specifying appropriate univariate, exponential family distributions for each dimension of x. The CMM takes the following form:

$$f(x) \quad \sim \quad \sum_{k=1}^{K} \pi_k \prod_{i=1}^{|x|} p_i(x_i|\theta_{k,i}) \tag{1}$$

where x is an observation vector of dimensionality $|x|$, K is the number of mixture components or clusters in the model, p_i is the distribution of the ith dimension (e.g. univariate Gaussian), and $\theta_{k,i}$ are the model parameters for the ith dimension distribution in the kth mixture component. The model structure implies that, given a mixture component, the dimensions of the observation vector are independent of one another. This simplicity facilitates jointly modeling the different observation types, leading to analytically tractable conditional and marginal distributions. However, it is important to note that, despite the features being treated as independent, complex correlations among the observation dimensions can be recovered by mixing (or averaging) over more components.

2.3. Composite Mixture Model Fitting to KPNC Dataset

Fitting a composite mixture model first requires specification of the univariate distributions for each dimension of the full data table. We support exponential family distributions for each observation dimension that can be efficiently estimated in parallel by computation of sufficient statistics. For our analysis, categorical variables (including mortality status) were modeled with categorical distributions while age, LAPS2, and COPS2, and the summary statistics of patient vital signs were modeled with either normal, gamma, or exponential distributions where appropriate. To facilitate model estimation, we added a small amount of jitter to non-categorical features. Specifically, we added a uniform random sample in the range $-d/10$ to $d/10$ to each value, where d is the smallest distance between unique values of that feature.

To fit a model for a given number of clusters, we performed expectation-maximization [7]. As expectation-maximization is a local optimization approach sensitive to initialization, we performed 3 random restarts of each fitting procedure with a maximum of 100 iterations. In our case, this number of iterations was generally sufficient for adequate model fitting. We determined the optimal number of clusters by evaluating the Bayesian information criterion (BIC) of all fitting restarts for each number of clusters tested (from 10 to 1000 clusters in increments of 10 for the individual post-admission datasets) and selecting the model with the lowest BIC. We also considered alternative model selection criteria such as Akaike information criterion (AIC), though BIC generated more reasonable results considering its harsher penalty on model complexity. The optimal number of clusters for the 3h, 6h, and 12h datasets was 250, 220, and 200, respectively. All composite mixture model data structures and fitting routines were implemented in a custom R software package (available upon request).

2.4. Mortality Enrichment and Cluster Trajectory Analysis

In addition to identifying structure in hospitalization episodes that can facilitate their stratification and, ideally, more effective treatment, we were interested in identifying changes in the physiological status of a patient during a hospitalization. To this end, we combined the 3h, 6h, and 12h datasets into a single dataset (hereafter referred to as the *combined* dataset) and fit a CMM to all observations. In this way, we treat the three observations for each episode as independent, and while the static features were shared (driving all three observations to co-cluster with one another), changes in the vitals over the course of hospitalization drive association of a given episode with different clusters over time.

To generate cluster assignments for each episode and post-admission period, we fit a CMM to the combined dataset, resulting in an optimal number of 650 clusters (tested 50 to 1000 in increments of 50), with cluster sizes ranging from 88 to 2,883 observations. As this number of clusters was prohibitively large for visualization purposes, we clustered the estimated parameters of the fitted model using the partitioning around medoids (PAM) algorithm [13] with the final number of clusters set to 20. We then mapped the newly assigned cluster labels to each observation in the dataset. We used these cluster labels to evaluate the number (and proportion) of episodes assigned to each cluster as well as the marginal mortality enrichment of each cluster as compared to the overall population for each of the three post-admission periods. Mortality enrichments were computed by a one-tailed Fisher's exact test.

To analyze temporal patterns of cluster membership, we combined the cluster labels for each of an episode's three post-admission periods into a single vector (*cluster trajectory*). This analysis resulted in 1,993 unique cluster trajectories. We then computed the significance of enrichment of each cluster trajectory for mortality events using a one-tailed Fisher's exact test. Adjusting for multiple hypothesis testing with a Bonferroni-corrected significance level of $\sim$ 2.51E-05, we identified 18 trajectories significantly enriched for mortality events when compared with the number of deaths in the full dataset (N=3,448).

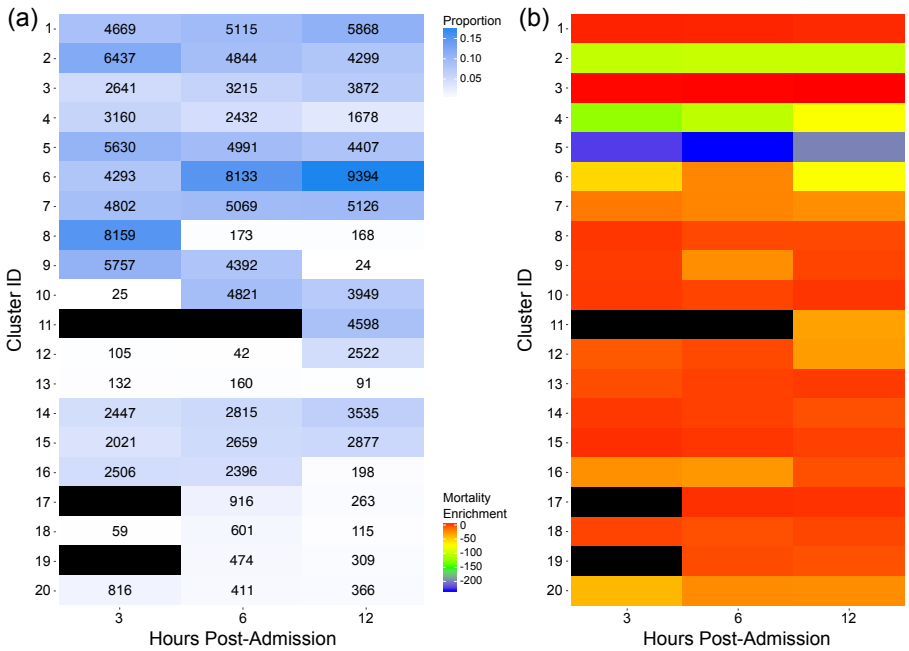


Figure 1: (a) Proportions of episodes assigned to each of the final 20 clusters for each post-admission period. The corresponding number of episodes assigned to each cluster are shown in each cell. (b) Mortality enrichment for each cluster and each post-admission period. The enrichment value is $\mathsf{log}(\text{-}\mathsf{log}(p))$ (for exposition purposes) where p is a one-tailed Fisher's exact test p-value evaluating the significance of enrichment of mortality events for each cluster during a given post-admission period. Black cells indicate clusters to which no episodes were assigned during that post-admission period.

To evaluate the physiological signatures of the three mortality-enriched clusters at 12h post-admission (clusters 1, 3, and 10), we first normalized the LAPS2, COPS2, and vital sign features to values between 0 and 1 by performing feature scaling (i.e. subtracting each feature by its minimum and dividing by the difference between the feature's maximum and minimum). We then conducted Wilcoxon rank-sum tests, comparing each cluster-specific distribution of LAPS2, COPS2, and vital sign statistics with their respective overall population distributions and computing 95% confidence intervals on the estimated difference between the cluster and overall population.

2.5. CMM-Derived Marginal Importance Analysis

Marginal importance plots are often used in conjunction with random forest models to visually evaluate the marginal effect of different features on the rate of a particular outcome [3, 14]. We adapted this concept to develop custom visualizations based on CMMs. We first determined the optimal number of clusters for a dataset using the BIC criterion as described in Section 2.3. We then computed the conditional posterior probabilities of cluster membership of each

full observation vector in the dataset. If an unclustered observation vector has a weight of 1, then these conditional probabilities can be considered the partial weights of each full observation corresponding to their probability of belonging to each CMM cluster. More formally, if we consider the K-dimensional vector, w_i, as the partial weights for each of K clusters of an episode i, then:

$$W_{i,j} = \mathsf{Pr}(Z_i = j | x_i, \hat{\theta}_{\mathsf{CMM}}) \tag{2}$$

Here, Z_i is the cluster label and x_i is the full observation vector, respectively, for episode i. $\hat{\theta}_{\mathsf{CMM}}$ are the estimated CMM model parameters. We stack these partial weight vectors for all episodes together (row-wise) to create the $N \times K$ matrix, W. We then enumerate equally spaced points in the range of each vital sign feature of interest. At each of these points, we compute the cluster membership probabilities of the point, u, ($\omega_{\mathsf{FEAT}}(u)$ as well as the expected rate of mortality, averaging over cluster-specific rates of mortality for the latter. The elements of the K-dimensional vector-valued function, $\omega_{\mathsf{FEAT}}(u)$, are defined as follows:

$$\omega_{\mathsf{FEAT},k}(u) = \mathsf{Pr}(Z = k | U_{i(\mathsf{FEAT})} = u, \hat{\theta}_{\mathsf{CMM}}) \tag{3}$$

Here, we treat U as a dummy feature vector with all entries but that corresponding to the vital sign feature of interest (i(FEAT)) missing. Z is an indicator variable for the dummy feature vector's cluster label and $\hat{\theta}_{\mathsf{CMM}}$ is again the estimated CMM model parameters. In this way, the cluster membership probabilities ($\omega_{\mathsf{FEAT},k}(u)$) are computed at a particular value (u) of the vital sign feature of interest by averaging over the other missing dimensions of the dummy feature vector, U. We then generated 95% Wilson score intervals [30] at each point in the vital sign feature range by treating the CMM-based expected rate of mortality at each point u as $\hat{p}$. To compute the sample size parameter at a particular vital sign feature point, u ($\hat{n}_{\mathsf{FEAT}}(u)$), for the confidence interval estimate, we used the cluster membership probabilities ($\omega_{\mathsf{FEAT}}(u)$) at that point and the partial weight matrix of the full observations (W) as follows:

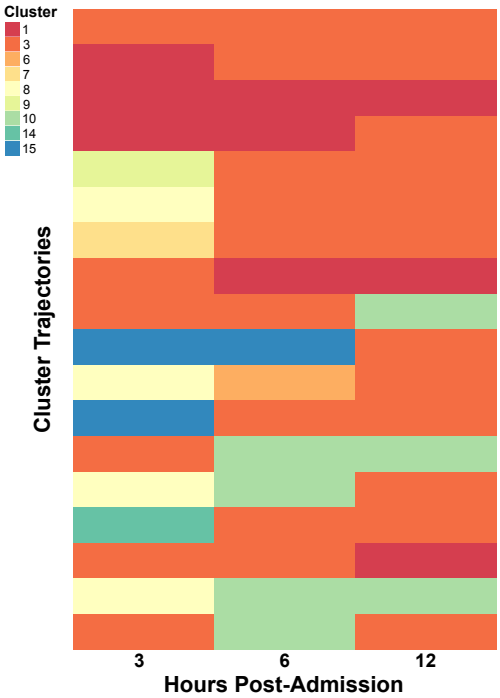
$$\hat{n}_{\mathsf{FEAT}}(u) = \sum_{i=1}^{N} [W \omega_{\mathsf{FEAT}}(u)]_i \tag{4}$$

Here, the matrix-vector product $W\omega_{\mathsf{FEAT}}(u)$ resulted in a N-dimensional vector of which we compute the sum. In this way, we computed the density (approximately the number of episodes) represented at a given vital sign feature point across all observations.

2.6. CMM-Based Missing Data Imputation Performance

To evaluate the CMM's missing data imputation performance, we introduced missing (completely at random; MCAR) data at a rate of 5% in the 6h dataset for all features except age, gender, and mortality, which were assumed non-missing for all records. The resulting dataset contained 12,344 complete records (i.e. records with no missing values).

Imputation using the CMM involves three steps: 1) performing a CMM fit on either A) the remaining complete records of the MCAR dataset or B) the MCAR dataset after performing population mean imputation, 2) re-computing the cluster membership probabilities for each episode conditioned on the non-missing features, and 3) imputing each missing value with either A) the cluster-averaged expected value (for non-categorical variables) or b) the label associated with the mode of the cluster-averaged expected category probabilities (for categorical features).



-log p-value	no. of deaths	total	mortality rate
876.18	717	1848	0.39
65.03	66	210	0.31
54.13	229	1782	0.13
45.90	45	142	0.32
45.67	47	155	0.30
32.36	46	202	0.23
24.17	21	64	0.33
23.11	31	135	0.23
21.21	13	29	0.45
19.58	19	66	0.29
18.45	19	70	0.27
16.71	20	84	0.24
16.55	8	15	0.53
15.77	19	81	0.23
14.52	19	87	0.22
11.97	17	85	0.20
11.67	122	1289	0.09
11.35	5	10	0.50

Figure 2: Cluster trajectories in the KPNC sepsis cohort enriched for mortality events. Trajectories are ranked in descending order by the -log p-value of a one-tailed Fisher's exact test. A cluster trajectory appears in each row and consists of three cluster assignments, one for each of the three post-admission periods (3h,6h,12h). For example, the top cluster trajectory indicates episodes that were assigned to cluster 3 at 3, 6, and 12 hours post-admission.

We compared imputation using the CMM to population mean imputation (as a baseline) and multivariate imputation using chained equations (MICE), a gold standard for data imputation in health sciences [4, 27]. The MICE procedure involves several steps: 1) imputing "placeholder" values with population mean imputation, 2) computing a regression for each feature column with missing values, treating all other features as dependent variables and using only records with an observed (i.e. non-placeholder) value of that feature, 3) replacing the placeholder value for each feature with the value predicted using the

corresponding regression, 4) repeating Steps 1 - 3, using the imputed values as new placeholder values, until the coefficients of the regression models converge. The prediction method used in Step 3 was predictive mean matching [4] for non-categorical variables and logistic/polytomous regression for categorical variables.

For each feature column with missing values, imputation performance was characterized by a distance, D. For non-categorical features, we used the sum of squared deviations from the observed values: $D = \sum_i^N (\hat{x}_i - x_i)^2$, where x_i and $\hat{x}_i$ are the observed and imputed values for the ith imputed record, respectively, and N is the total number of imputed records. For categorical features, we used the counts of incorrectly imputed values: $D = \sum_i^N [\hat{x}_i \neq x_i]$, where brackets indicate 1 when the condition is true and 0 otherwise. Distances for each feature were then normalized by the corresponding distance obtained using population mean imputation. We note that most choices for distance metric for categorical features (e.g. overlap, occurrence frequency, inverse occurrence frequency, Eskin [8], Goodall [9]) produce identical distances after normalization.

3. Results

3.1. Mortality Enrichment and Physiological Signatures of CMM Clusters

An important challenge in treatment of sepsis is the detection of (sometimes subtle) physiological changes in a patient while accounting for the heterogeneous presentation of the condition. We characterize these latent changes by modeling the full demographic and physiological space of all episodes, regardless of time, subsequently identifying the regions or clusters of this space with which a patient episode associated during a hospitalization. To this end, we concatenated the 3h, 6h, and 12h datasets into a *combined* dataset, fit a CMM to all observations, and generated assignments for each episode at each post-admission time to one of twenty clusters. Observations from episodes in which patient physiology did not dramatically change over time would co-cluster while episodes marked by physiological changes would be assigned to different clusters over time.

From this analysis, we determined the proportion of episodes in each cluster at each post-admission time and identified those clusters more or less associated with mortality events. We found that while the majority of the episodes were assigned to one of the top seven clusters over all post-admission periods, some clusters were associated with progressively fewer episodes from 3h to 12h post-admission (e.g. cluster 8) while other clusters (e.g. cluster 10) became associated with more episodes over time (Figure 1a). These patterns highlight the diverse demographic/physiological "landscape" through which patients progressed during their hospitalization and indicate how some clusters represent transient "stops" in a patient's progression. When we considered each cluster separately, we found that mortality events were significantly underrepresented in episodes assigned to clusters 2, 4, and 5. Conversely, clusters 1, 3, 10, and 17 showed high degrees of enrichment for mortality events (Figure 1b), indicating population heterogeneity among those patient episodes considered at high risk for mortality.

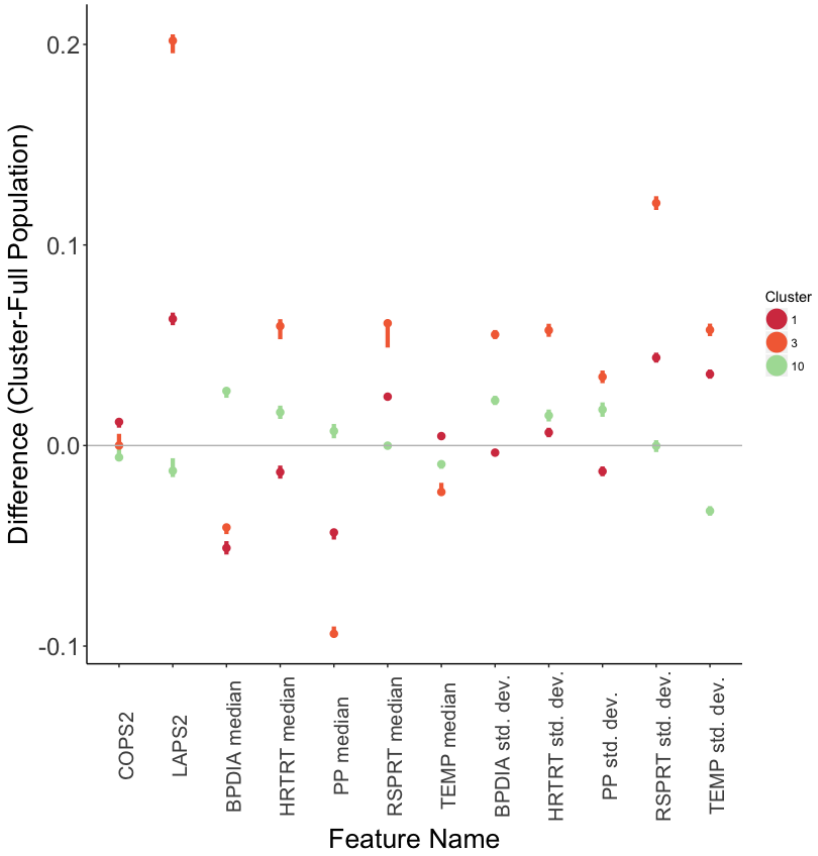


Figure 3: Estimated differences between cluster-specific and population values of each feature. The dots represent the estimated difference while whiskers represent the 95% confidence interval as computed by Wilcoxon rank-sum tests. The colors of each cluster's dot and whiskers correspond to those in Figure 2. The lack of an overlap between these intervals and a difference of 0 (gray line) reflects a significant difference in cluster-specific values of the feature as compared to the overall population.

To further analyze temporal patterns in cluster membership, we also considered the full sequence of cluster assignments for each episode's observations (*cluster trajectory*). This analysis resulted in 1,993 unique cluster trajectories out of the 53,659 episodes, 18 of which were highly enriched for mortality events (Figure 2). The trajectory most enriched for mortality events involved episodes in which the patient remained in cluster 3. Indeed, all but two of the trajectories enriched for mortality events associated with cluster 3 at some point during the hospitalization. For mortality-enriched trajectories, episodes tended to associate with a diverse set of clusters at 3 hours post-admission but generally transitioned into one of three clusters (1, 3, or 10) at 12 hours post-admission.

To assess physiological signatures of these three clusters at 12 hours post-

admission, we evaluated the statistical significance of the differences between the distributions of cluster-specific and overall population features (Figure 3; described in Methods). Our analysis indicates that while all three clusters associated with an elevated mortality risk, they reflect distinctly different physiological populations, highlighting the heterogeneity inherent in the presentation of sepsis. For example, while cluster 3 episodes showed significantly higher acute disease burden (LAPS2) than those episodes in clusters 1 and 10, cluster 1 episodes showed significantly higher chronic disease burden (COPS2). Interestingly, cluster 10 episodes showed significantly less chronic and acute disease burden than the overall population.

In addition, the vital sign characteristics of the different mortality-enriched cluster episodes were distinctly different. Vital sign median levels and standard deviations for cluster 3 generally deviated more significantly from those in the overall dataset when compared with the same characteristics in clusters 1 and 10. In particular, cluster 3 episodes showed significantly lower median blood pressure, pulse pressure, and body temperature, higher median heart and respiratory rates, and the highest vital sign variability. Cluster 1 vital sign patterns were generally similar to those of cluster 3. However, cluster 1 episodes did show significantly lower heart rate, higher body temperature, and generally lower vital sign variability for blood pressure, heart rate and pulse pressure. Cluster 10 episodes also associated with different physiological trends: higher median blood pressure and pulse pressure, lower variability in body temperature, and respiratory rate median levels and variability close to those of the overall population. This analysis suggests that while a patient may be placed in a high-risk strata, their physiological state can be quite distinct from that of another high-risk patient. These patterns help provide physiological insights for the treatment of patients as they transition through these states over the course of a hospitalization.

3.2. Evaluating Marginal Physiological Associations with Mortality Rate

The CMM enables stratification of populations into physiologically distinct subgroups that could highlight the need for different treatment regimens. The CMM highlighted the values of physiological features (e.g. median blood pressure) that demonstrated greater marginal association with mortality events. We displayed these patterns by adapting a technique often used in conjunction with random forest models: marginal importance plots [3, 14]. These plots show the expected mortality rate under a fitted CMM, along with 95% confidence intervals, at different values in the range of a given vital sign feature of interest (Figures 4a, 4b, 4c, and 4d). For this analysis, we fit the CMM model to the 12h dataset only. Trends in these visualizations reflect current clinical knowledge and practices: median diastolic blood pressure and body temperature levels associated with low levels of mortality at 12 hours post-admission are within a range (i.e. "within normal limits") outside of which the mortality rate begins to increase.

In the case of median diastolic blood pressure (Figure 4a), the mortality rate varied between ∼6% (similar to the overall population rate) at values near 75 to

rates in excess of 50% at values near 25, reflecting the importance of blood pressure management in sepsis ([10] and references therein). Likewise, in the case of body temperature (Figure 4c), mortality rates were at their lowest (∼5%) at approximately 98°F, climbing to rates of nearly 60% as body temperature dropped to 85°F. In the case of standard deviations of the different vital sign features (Figures 4b and 4d), the plots indicate that, in general, increased variability in a patient's vital signs is associated with higher rates of mortality. However, this relationship was not linear (e.g. Figure 4d), suggesting that some variability was not harmful for patients, within a specific range dependent on each variable. Overall, in stratifying the population, the CMM helps to establish fine-grained boundaries of demographic/physiological space that are strongly associated with increased mortality, thus facilitating more precise characterization of the marginal associations of physiological variables with increased rates of mortality or other outcomes of interest.

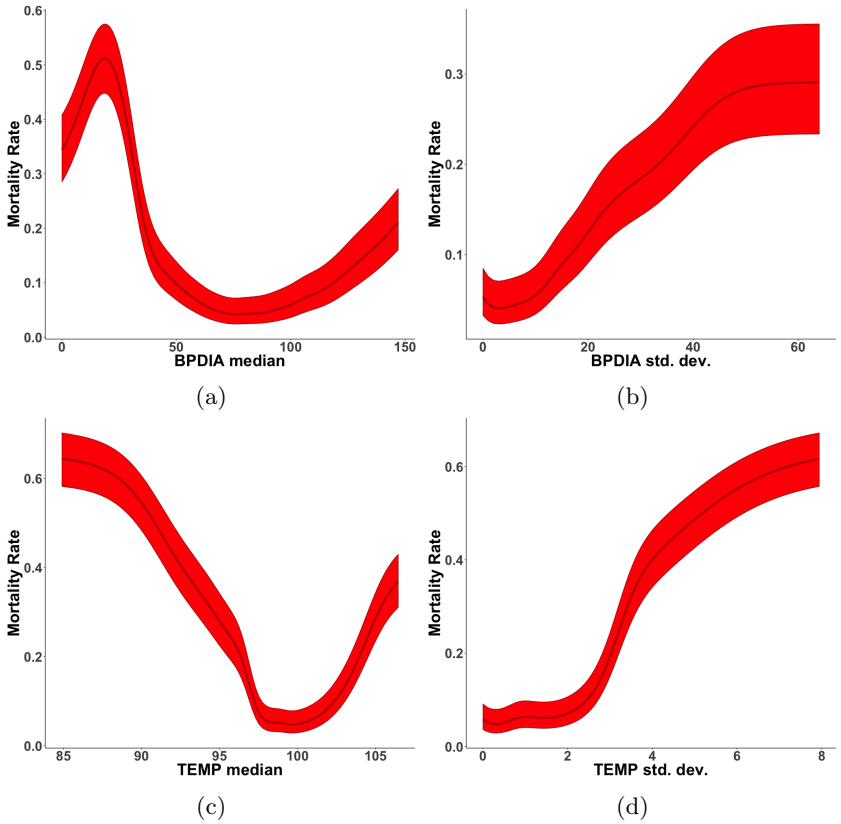


Figure 4: Marginal importance plots for diastolic blood pressure (median (a) and standard deviation (b)) and pulse pressure (median (c) and standard deviation (d)) at 12 hours post-admission. The dark line in the center of the band is the estimated mortality rate at each value of the vitals feature of interest while the lighter bands are 95% Wilson's score intervals.

3.3. CMM-Based Missing Data Imputation Performance

As the CMM is a joint probability model, we can use the learned structure from a trained CMM to impute missing observations, a common task in EMR analysis. Indeed, recent unsupervised structure learning approaches have proven successful at this task [2]. We compare the CMM's performance on a modified version of the 6h dataset (with 5% missing data introduced; see Methods) with that of two other imputation approaches: 1) MICE [4], a gold standard method for missing data imputation and 2) imputation with the population mean (our baseline). As MICE (based on fully conditional regressions) and CMMs (based on unsupervised learning) are designed for very different purposes, our goal in evaluating the CMM's data imputation performance is to not to draw direct comparisons to stand-alone imputation methods, but rather to assess the situations in which CMMs might prove useful for data imputation. We tested two variants of our CMM-based approach in which we either trained the model with the full dataset after imputing with each feature column's population mean or solely using complete records. Figure 5 shows the distances, relative to population mean imputation, for select features using various imputation methods.

For all features, CMM imputation using estimates based on population mean imputation outperformed CMM imputation using estimates based solely on complete records (Figure 5). This result was expected, as the former has over four times as much training data as the latter. For all summary statistics of vital signs, MICE outperformed both CMM imputation methods. As MICE leverages fully conditional data in its imputation and vitals are expected to be globally predictive of one another, this result is not surprising. No approach significantly improved over population imputation for the three categorical features in the dataset (membership status, transport code, and facility code). In fact, the variant of MICE we used intentionally introduces variability in its predictions, aiming to faithfully recapitulate overall variability in a feature dimension with missing data as opposed to optimally impute the missing points. We attribute the result we observe for the continuous physiological feature, COPS2, to this characteristic of MICE. Moreover, in the case of CMM-based imputation, we don't expect the population to be well-structured with respect to the categorical features which are largely operational.

CMM-based imputation did result in substantially smaller distances for one important physiological feature: LAPS2. While LAPS2—which is based on laboratory information obtained within the 24 hours preceding hospitalization—shows very little conditional dependence on the other features in the dataset across the entire population (hence the slight improvement MICE yields over population mean imputation), the CMM seems to detect latent structure according to this measure of acute illness, offering a significant improvement over population mean imputation. This result corroborates our previous findings of population structure according to mortality risk. Overall, the average relative distances across all features are 0.570 ± 0.001 (CMM using complete records), 0.524 ± 0.001 (CMM using population mean imputation), and 0.398 ± 0.009 (MICE), indicating the CMM's general utility for missing data imputation.

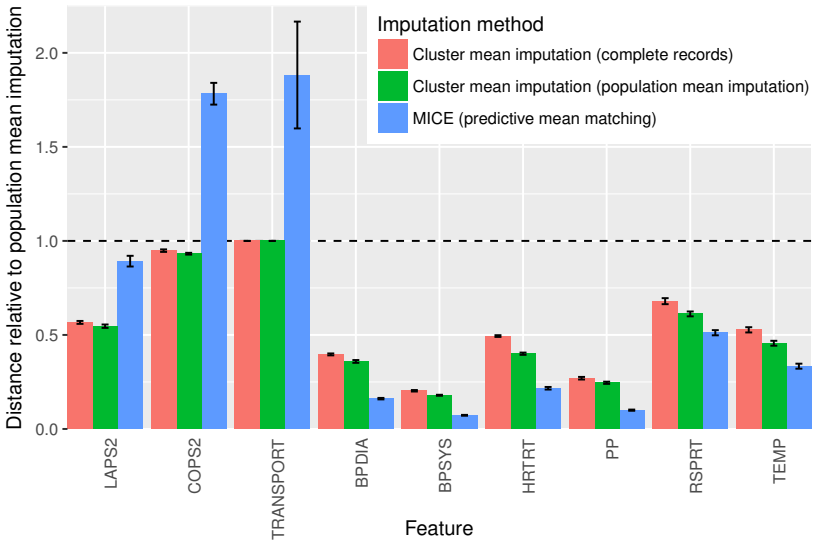


Figure 5: Distances between imputed and observed values from three different imputation methods for select features, relative to distances from population mean imputation ($D_{\text{normalized}} = D_{\text{imp.}}/D_{\text{pop. mean imp.}}$), where D is the appropriate distance for each feature as described in section 2.6). Error bars represent standard deviation across imputations on 3 independent CMM fits (for CMM imputations) or across 20 independent imputations (for MICE). Performance was similar for TRANSPORT and two other categorical features (membership and facility codes) not shown. TRANSPORT, patient transport code; BPDIA, median diastolic blood pressure; BPSYS, median systolic blood pressure; HRTRT, median heart rate; PP, median pulse pressure; RSPRT, median respiratory rate; TEMP, median temperature.

4. Discussion & Conclusions

We demonstrated the flexibility and efficacy of our composite mixture model approach in analyzing multi-typed EMR datasets characterized by population heterogeneity and data missingness. The model can be used to stratify clinical populations according to risk of different unfavorable outcomes at different points in a hospitalization and to evaluate the marginal associations of patient physiology with higher rates of such outcomes. As a joint rather than conditional probability model, the CMM also provides a useful framework to both impute missing observations and predict outcomes of interest for held-out samples, enabling analyses that can complement purely predictive modeling studies.

The CMM is not without its drawbacks. For one, manually specifying the model template, or list of univariate distributions that model each feature column of a dataset, can be prohibitive in high feature dimensions. Also, while model fitting routines have been vectorized, they were not optimized further. As such, estimation of a CMM can be computationally intensive (e.g. approximately 3.5 hours on a single core for the 3 hr dataset, with 53,659 episodes and 32 features, fitting to 250 (the optimal number) clusters). In previous non-medical applications, we overcame this computational burden by leverag-

ing scalable high-performance computing software framework for streaming data analysis [26]. We are currently extending our model fitting software to automatically identify "default" distributions for a given feature column as well as exploring scalable (i.e. data-parallel) estimation routines.

In addition to demographic information, we opted to model a particular set of patient vital sign statistics representative of level and variability over fixed post-admission periods. However, this approach ignores dynamic patterns in the vital signs including changes in the cross-correlation between different vital sign types that might also be indicative of changes in patient health [18] and could be included as features. In addition, an alternative approach to modeling the vital signs would be to extend the CMM with distributions appropriate for time series data (e.g. a Gaussian process; [21]). Such an extension could potentially reduce the number of parameters in the overall CMM and lessen model specification and estimation time. Moreover, while not included in our analyses, lab test results can aid in the risk stratification of patients and, like vital signs, provide clinically actionable targets for intervention to improve patient health. As part of our ongoing and future work, we are investigating these extensions to our analyses.

We achieved competitive missing data imputation performance with the CMM, though it is important to note how that analysis also highlights differences between unsupervised and supervised learning approaches. Regression-based methods like MICE make predictions based on population-wide conditional information. Thus, features for which MICE performs well could indicate strong conditional dependencies between those features and all other features at the population level. In contrast, CMM-based imputation makes predictions based on identifying latent local structure, identifying subpopulations of the full dataset that are more similar to one another. In particular, the CMM uncovers latent structure with respect to features like LAPS2, leading to improved imputation performance, even though the current formulation of the model does not include conditional dependencies. It is worth noting that our assumption of conditional independence among feature dimensions given a mixture component represent one potential model structure of the CMM. Conditional dependencies can be introduced into the CMM rather straightforwardly (e.g. replacing univariate normal distributions with linear regressions conditioned on other features) and will be the subject of future work. Taken together, these results suggest that the CMM could be used in concert with other data imputation approaches to address missing data challenges on a per-feature basis.

Our CMM-based analysis of the KPNC sepsis cohort revealed physiologically distinct subpopulations associated with elevated rates of mortality. Those episodes most enriched for mortality events tended to be assigned to cluster 3 at 12 hours post-admission (Figure 2). However, an important follow-up analysis would identify those episodes assigned to high-mortality clusters at 3 hours post-admission that were assigned to low mortality clusters by 12 hours post-admission. Importantly, as our full dataset also includes discrete time series of the medications and procedures ordered by the clinician during each hospitalization, we can evaluate whether certain courses of action were over-represented

in trajectories in which patients transitioned out of high mortality clusters and into low mortality clusters. Overall, our flexible CMM framework provides a valuable decision support tool to characterize and stratify heterogeneous clinical populations, capturing clinically relevant changes in patient physiology while accounting for the data missingness and wide variety of data types inherent in large-scale electronic medical record data.

5. Acknowledgements

The authors would like to thank Priyadip Ray for critical reading of the manuscript. Dr. Vincent X. Liu is funded by NIH grant (NIH K23GM112018). This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344 (LLNL-JRNL-730845-DRAFT).

6. References

[1] Brett K. Beaulieu-Jones and Casey S. Greene. Semi-supervised learning of the electronic health record for phenotype stratification. *Journal of Biomedical Informatics*, 64:168 – 178, 2016.

[2] Brett K. Beaulieu-Jones, Jason H. Moore, and The Pooled Resource Open-Access ALS Clinical Trials Consortium. Missing Data Imputation in the Electronic Health Record using Deeply Learned autoencoders. In *Biocomputing 2017*, pages 207–218.

[3] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] Stef Buuren and Karin Groothuis-Oudshoorn. MICE: Multivariate imputation by chained equations in R. *Journal of statistical software*, 45(3), 2011.

[5] Robert J Carroll, Anne E Eyler, and Joshua C Denny. Intelligent use and clinical benefits of electronic health records in rheumatoid arthritis. *Expert Review of Clinical Immunology*, 11(3):329–337, 2015.

[6] Seymour CW, Liu VX, Iwashyna TJ, and et al. Assessment of clinical criteria for sepsis: For the third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):762–774, 2016.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[8] Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy, and Sal Stolfo. A geometric framework for unsupervised anomaly detection. In *Applications of data mining in computer security*, pages 77–101. Springer, 2002.

[9] David W Goodall. A new similarity index based on probability. *Biometrics*, pages 882–907, 1966.

[10] Jeffrey E Gotts and Michael A Matthay. Sepsis: pathophysiology and clinical management. *BMJ*, 353, 2016.

[11] Douglas Hanahan and Robert A. Weinberg. Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674.

[12] Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. A targeted real-time early warning score (TREWScore) for septic shock. *Science Translational Medicine*, 7(299):299ra122–299ra122, 2015.

[13] Leonard Kaufman and Peter J. Rousseeuw. Partitioning Around Medoids (Program PAM). In *Finding Groups in Data*, pages 68–125. John Wiley & Sons, Inc., 2008.

[14] M. Kuhn and K. Johnson. *Applied Predictive Modeling*. SpringerLink : Bücher. Springer New York, 2013.

[15] Thomas A. Lasko, Joshua C. Denny, and Mia A. Levy. Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data. *PLOS ONE*, 8(6):1–13, 06 2013.

[16] John C. Marshall. The staging of sepsis: understanding heterogeneity in treatment efficacy. *Critical Care*, 9(6):626, 2005.

[17] Riccardo Miotto, Li Li, Brian A. Kidd, and Joel T. Dudley. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*, 6:26094 EP –, 05 2016.

[18] J Randall Moorman, John B Delos, Abigail A Flower, Hanqing Cao, Boris P Kovatchev, Joshua S Richman, and Douglas E Lake. Cardiovascular oscillations at the bedside: early diagnosis of neonatal sepsis using heart rate characteristics monitoring. *Physiological Measurement*, 32(11):1821–1832, 11 2011.

[19] Chris Paxton, Alexandru Niculescu-Mizil, and Suchi Saria. Developing Predictive Models Using Electronic Medical Records: Challenges and Pitfalls. In *AMIA 2013, American Medical Informatics Association Annual Symposium*, 2013.

[20] Rimma Pivovarov, Adler J. Perotte, Edouard Grave, John Angiolillo, Chris H. Wiggins, and Nomie Elhadad. Learning Probabilistic Phenotypes from Heterogeneous EHR Data. *Journal of Biomedical Informatics*, 58(C):156–165, December 2015.

[21] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[22] A. P. Sales, C. Challis, R. Prenger, and D. Merl. Semi-supervised classification of texts using particle learning for probabilistic automata. In *Bayesian Theory and Applications*. Oxford University Press, 2013.

[23] Peter Schlattmann. *Medical Applications of Finite Mixture Models*, volume 1. Springer, 2009.

[24] Stuart J Schnitt. Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy. *Mod Pathol*, 23(S2):S60–S64, May 2010.

[25] Liu V, Escobar GJ, Greene JD, and et al. Hospital deaths in patients with sepsis from 2 independent cohorts. *JAMA*, 312(1):90–92, 2014.

[26] T. Wasson and A. P. Sales. Application-Agnostic Streaming Bayesian Inference via Apache Storm. In *International Conference on Big Data Analytics*, 2014.

[27] Brian J Wells, Kevin M Chagin, Amy S Nowacki, and Michael W Kattan. Strategies for handling missing data in electronic health record derived data. *eGEMs*, 1(3), 2013.

[28] Jenna Wiens, Wayne N Campbell, Ella S Franklin, John V Guttag, and Eric Horvitz. Learning Data-Driven Patient Risk Stratification Models for *Clostridium difficile*. *Open Forum Infect Dis.*, 1(2):ofu045, 2014.

[29] Jenna Wiens, John Guttag, and Eric Horvitz. Patient Risk Stratification with Time-varying Parameters: A Multitask Learning Approach. *J. Mach. Learn. Res.*, 17(1):2797–2819, January 2016.

[30] Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22:209 – 212, 1927.