Wave-MixEHR: Topic Modeling of Electronic Health Records with ECG Waveforms

Payal Chandak CHANDAK@MIT.EDU
Daniel Shao DSHAO@MIT.EDU
Joanna Shen JOANNASH@MIT.EDU
Morgan Talbot MTALBOT@MIT.EDU

Abstract

Electronic health records (EHRs) are a valuable source of patient health information that can facilitate disease-risk prediction, precision medicine, and actionable clinical recommendations. However, the heterogeneity of data types and sources can make it difficult to combine and analyze the data effectively. MixEHR is a multi-view Bayesian topic model that addresses these challenges by using clinical notes, lab results, and prescription data to uncover meaningful combinations of clinical features and produces topic mixtures for each patient. However, no topic model has yet incorporated electrocardiogram (ECG) signals, despite the rich, complex, and clinically relevant information they contain. In this work, we introduce Wave-MixEHR, a multi-modal Bayesian topic model that incorporates ECG signals, clinical notes, lab tests, and prescription information. We demonstrate the effectiveness of our model at generating interpretable and clinical relevant topics on the MIMIC-III Waveforms dataset. We demonstrate a variety of multimodal, clinically relevant disease phenotype associations, including links between cataracts, hyperuricemia, and atrial fibrillation that were also established in recent large cohort studies. We also identify a potentially novel association between bile duct obstruction and parasympathetic tone as measured by ECG signals, which is well-supported by rodent studies but, to our knowledge, is yet to be characterized in humans. We also evaluate the performance of supervised learning models that use the topic distributions of each patient from Wave-MixEHR's output to predict mortality and length of hospital stay. We experiment with various data fusion methods, including early, intermediate, and late fusion, and show that ECG signals, when analyzed in conjunction with other clinical modalities, provide valuable context for predicting patient outcomes. Our proposed approach, Wave-MixEHR, confers competitive performance and reveals meaningful disease-related topics, offering a promising avenue for improving clinical decision-making. Our source code is publicly available on GitHub.

1. Introduction

Electronic Health Records (EHRs) are vast, heterogeneous repositories of patient health information that offer unique opportunities for health informatics, disease-risk prediction, clinical decision-making, and precision medicine. Nonetheless, they also pose substantial modeling obstacles, such as highly sparse data matrices, noise-ridden clinical notes, arbitrary billing code assignment biases, diagnosis-driven lab tests, and a diversity of data types including information-rich waveforms and medical images.

MixEHR (Li et al., 2020) is a multi-view Bayesian topic model designed to tackle the challenges inherent in analyzing EHRs. To circumvent these issues, MixEHR employs a probabilistic joint matrix factorization strategy. This approach projects each patient's high-

dimensional and diverse clinical record onto a low-dimensional probabilistic meta-phenotype signature, thereby encapsulating the patient's mixed memberships across various latent disease topics. MixEHR's novelty lies in its two-tiered factorization process, which enables it to manage highly heterogeneous data types effectively. At a more granular level, it applies data-type-specific topic models, deriving a set of basis matrices for each data type. For instance, in the context of the MIMIC-III dataset (Johnson et al., 2016; Goldberger et al., 2000), MixEHR successfully discerned seven basis matrices corresponding to clinical notes, ICD-9 billing codes, prescriptions, DRG billing codes, CPT procedural codes, lab tests, and lab results. While MixEHR reveals meaningful disease-related topics on these patients, it does not leverage all modalities of the data available from the MIMIC-III dataset.

Patients admitted to the Intensive Care Unit (ICU) are typically connected to bedside telemetry monitors, which persistently record a wide array of waveforms. These high-frequency waveforms capture diverse measurements such as electrocardiogram (ECG), respiration, blood pressure, along with other vital signs. ECGs trace the electrical activity of the heart, thereby serving as invaluable tools in discerning cardiac ailments and monitoring heart function. The constant acquisition of high-resolution ECG data presents a distinct opportunity to deepen our understanding of these patients' disease conditions. By leveraging the wealth of information contained within ECGs, we seek to identify topics that are particularly relevant to cardiac disease and mortality, including multimodal associations that might otherwise be missed, and to develop a more holistic representation of each patient's status.

Here, we augment MixEHR into **Wave-MixEHR**, by integrating ECG data into the Bayesian topic model. We use the MIMIC-III Waveforms dataset, generate categorical features from the continuous ECG waveforms, and incorporate them into MixEHR's basis matrices. Wave-MixEHR not only improved accuracy on downstream classification tasks, but also generated clinically meaningful topics with interpretable associations between ECG features and other modalities of clinical information. Wave-MixEHR can ingest a comprehensive snapshot of ICU patients with the rich information of ECG waveforms, and thereby offer intriguing topic models and enhanced performance. This approach could yield new discoveries and decision support applications that pave the way for better-informed treatment decisions.

2. Related Work

Our work builds on existing research in clinical multimodal topic modeling, machine learning for mortality prediction, and topic modeling on ECG data. In the field of machine learning and healthcare, several previous studies have focused on combining structured and unstructured clinical data to leverage the wealth of information contained in EHRs.

Topic modeling has emerged as a valuable approach for extracting meaningful topics from unstructured EHRs. Several studies have investigated the effectiveness of various topic modeling algorithms, including Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF), in identifying clinically relevant topics from extensive EHR data. Halgrim et al. (2011) conducted a comparative analysis of machine learning techniques for extracting medical knowledge from clinical text (Halgrim et al., 2011), while Luo et al. (2017) proposed a practical methodology for clinical topic modeling in EHRs (Luo

et al., 2017). Other research endeavors have focused on identifying medical terms and clinical topics by combining rule-based and machine learning methods, as demonstrated by Cuzzola et al. (2019)(Jovanovic et al., 2014). Furthermore, integrating clinical data into topic models to enhance cohort identification and characterization has been explored by Ghassemi et al. (2014) (Marzyeh Ghassemi and Szolovits). By leveraging topic modeling techniques in the clinical domain, researchers aim to uncover valuable insights, improve clinical decision-making, and advance healthcare research and practice.

One especially promising recent approach is MixEHR, a state-of-art method for clinical topic modeling that has shown promising results for discovering clinically meaningful latent topics from EHRs. MixEHR topics can also be used to impute missing data and perform classification tasks, such as predicting ICU mortality (Li and Kellis, 2018). Some recent work has extended the MixEHR approach to align the discovered topics with known phenotypes (Song et al., 2021) and infer the contributions of different clinical specialties to the care each patient needs (Ahuja et al., 2022).

Existing topic modeling approaches still have limitations. MixEHR models are multimodal, leveraging data from unstructured clinical notes, ICD-9 codes, prescriptions, lab tests, current procedural terminology, and diagnosis-related groups. However, the existing MixEHR approaches do not make direct use of ECG data, which may contain rich information relevant to many healthcare and clinical machine learning tasks. Although topic modeling approaches have been used on ECG data in isolation, including for detection of myocardial infarctions (Sun et al., 2012) and finding latent structure in patterns found in long-term ECG recordings (Van Esbroeck et al., 2021), we are unaware of any existing works that include ECG data in EHR topic modeling.

3. Methods

3.1. Generating Topics from Wave-MixEHR

We trained the original MixEHR model (Li et al., 2020) without any modifications to its source code, but with the addition of binarized ECG-derived features encoded similarly to ICD codes (see section 4). MixEHR learns to encode multimodal, high-dimensional, and extremely sparse EHR data from each patient or hospital admission into a low-dimensional space of latent "topics," where each patient has a mixed, weighted membership in all of the topics. Following the original MixEHR paper, we set the model to learn 75 topics. MixEHR builds on the same ideas used in Latent Dirichlet Allocation (LDA), a common topic modeling approach for natural language tasks. However, EHR data has multiple different modalities with very different dimensionality - for example, there are far more relevant keywords in clinical notes than features in most other modalities, so note keywords would tend to dominate LDA models that treat all features the same way. MixEHR addresses this problem by learning separate topic models for each modality before linking them into a multi-modal topic model using a loading matrix (Li et al., 2020). This approach makes MixEHR ideal for modeling our relatively small set of ECG features jointly with other EHR modalities of much higher dimensionality. We expanded MixEHR by developing a framework for incorporating continuous waveform signals, by converting them to categorical features based on distribution quantiles (see section 4 below).

MixEHR is a generative latent topic model that jointly maximizes the likelihood of EHR data in the training set. Referring to Fig. 1c in Li et al. (2020), we marginalize the tensors learned via collapsed variational Bayesian inference, which have patients, phenotypes (e.g., specific ICDs, lab tests, etc), and the 75 latent topics as their three dimensions. By averaging along the phenotype dimension, we estimate each patient's membership with each latent topic, which we then use as a feature set for supervised learning in section 5.2. By averaging along the patient dimension, we estimate the association of each phenotype with each latent topic, allowing us to interpret the latent topics in section 5.1.

3.2. Prediction of Clinical Outcomes

We hypothesized that the generated topics would serve as rich, clinically-relevant encodings of the multimodal inputs into Wave-MixEHR. To quantify the clinical relevance of the generated topics, we treated Wave-MixEHR's topics as features for an XGBoost model to predict 1) patient mortality during the hospital admission and 2) whether length of stay (LOS) would exceed 14 days. Additionally, we evaluated how different fusion methods for incorporating ECG features into our XGBoost model affected model performance. This ablation analysis allows us to elucidate the specific contribution of topic modeling for processing ECG data.

3.2.1. Fusion approaches

Let $A \in [0,1]^{N\times a}$ be binary lab results over N hospital admissions, $B \in [0,1]^{N\times b}$ be the one-hot encoding of clinical notes, and $C \in [0,2]^{N\times 150}$ be the set of discretized ECG features across 3 quantiles. Let $f(x) \in \{0,1\}^N$ be the binary output of XGBoost, and $m(x) \in [0,1]^{N\times 75}$ the topic assignments across all 75 topics.

- **Early Fusion**: We directly feed in the ECG features as an additional modality for Wave-MixEHR, whereby the final set of predictions $Y_{early} \in [0,1]^N$ is given by $Y_{early} = f(m(A,B,C))$.
- **Intermediate Fusion**: We generate topics from the original MixEHR modalities A, B, and concatenate C into the XGBoost input, such that $Y_{mid} = f(C + m(A,B))$
- **Late Fusion**: We average the predictions between one XGBoost model trained on $m(A,B)$, and another on C, such that $Y_{late} = \dfrac{f_1(C) + f_2(m(A,B))}{2}$

For all experiments, XGBoost was initialized with 100 estimators and max depth 7.

4. Data and Experiment setup

4.1. Dataset

We used the MIMIC-III clinical database (Johnson et al., 2016; Goldberger et al., 2000) and its Waveform Database Matched Subset (Moody et al., 2020), which is comprised of physiologic waveforms obtained from bedside telemetry monitors of ICU patients linked to their MIMIC-III records. This dataset contains 10,282 distinct ICU patients who have at least one ECG recorded. Lead II of the ECG is the most prevalent waveform in this database and we focus exclusively on this waveform here.

4.2. Building trajectories of ECGs

The dataset comprised numerous hours of continuous ECG recordings for each admission, which made it impractical for state-of-the-art deep learning systems to analyze in its unprocessed form. To address this challenge, we divided each patient's visit into non-overlapping 12-hour intervals (termed "trajectories"), and for each hour within these trajectories we randomly selected a representative 10-second ECG signal. We characterized a trajectory as a sequence of 12 ECG waveforms, wherein each waveform represented a 10-second ECG from Lead II, sampled at a frequency of 125 Hz.

To maintain high data quality, we excluded admissions with fewer than one trajectory from our analysis. We implemented this exclusion criterion because it suggested that patients underwent inconsistent monitoring in the ICU during their admission, which likely resulted in their ECGs being contaminated by numerous artifacts and noise. As most admissions contained multiple 12-hour trajectories, and our goal was to predict patient outcome at the end of the hospital admission, we only used trajectories from within the first 24 hours of admission when training Wave-MixEHR for downstream prediction. Similarly, we excluded ICD codes entirely since they lack timestamps, and only used other features collected within 24 hours of admission. This enabled our model to make the earliest possible predictions for tasks such as length of stay without introducing data leakage by exposing data from the end of the stay; after applying this data processing and filtering methodology, the final dataset included 4,917 patients and 5,235 admissions, where each admission had one corresponding trajectory.

4.3. Feature Extraction from ECGs

The ECG is a well-researched signal in the realm of medical literature (Camm et al., 2009), and we leverage not only expert clinical knowledge but also traditional signal processing techniques to inform our choice of ECG features. We generate 150 structured features, which fall under the following overarching categories: Power ratios, energy entropy, higuchi fractal, beats, heart rate variability. These categories are described in detail in the Appendix.

Clinically, it is typical to average the features computed at each hour over a 12-hour trajectory. Consequently, each admission is associated with a single set of features. To accommodate MixEHR's requirement for categorical features rather than continuous values, we calculated the 25th and 75th percentile values for the overall training set distribution of each feature, resulting in "low" and "high" binary features for each continuous feature.

4.4. Preparing EHR data for MixEHR modeling

We prepared data from the MIMIC-III dataset (Johnson et al., 2016) by building a preprocessing pipeline similar to the original MixEHR paper (Li et al., 2020). We used the NLTK Python library (Loper and Bird, 2002) to generate a dictionary of keywords from clinical notes, removing common stop words. Notably, because our end-goal was to predict patient outcome at the end of the hospital admission, we only used data from the first 24 hours of admission. Unlike the original MixEHR paper, we also exclude ICD codes from the input because the time of diagnoses/procedures cannot be determined.

We encoded all selected features as an array of integers representing patient ID, modality ID, phenotype ID (e.g. for different lab tests), state ID (different results of lab tests, or

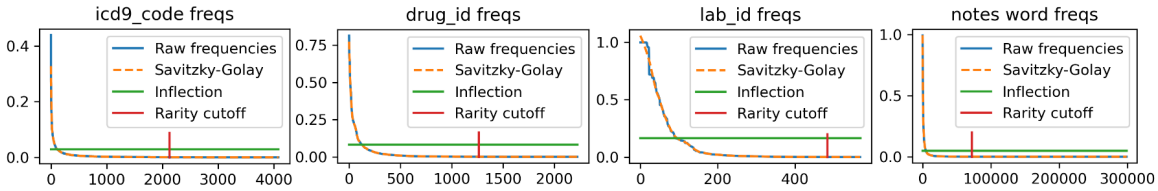

Figure 1: To remove features common to many patients, we generated feature frequency curves for each non-ECG modality. The horizontal axes show the number of features before filtering, and features are sorted by their frequency among patients (which is the vertical axis). We smoothed the frequency curves with a Savitzky-Golay filter before numerically calculating inflection points.

uniformly "1" for other modalities) and frequency (how many times each phenotype ID and state ID combination occurred for the patient). We encoded our binarized ECG quantile features similarly to prescriptions, and clinical note keywords, but with frequency always set to "1" because each patient either had a given feature in their ECG data or they did not (as opposed to, for example, multiple instances of the same word in a set of notes). We also did not remove common or rare ECG features as with other modalities, because our binarized ECG features are based on quantiles.

At the beginning of data preprocessing, we randomly assigned patients into a train/validation/test split with 80%, 10%, and 10% of the patients respectively. To avoid the potential problem of multiple admissions per patient possibly for different reasons, we randomly selected one admission per patient that fulfilled our ECG preprocessing criteria: this resulted in a less than 6% data loss because most patients only had one admission. In all of our preprocessing steps, we relied exclusively on information from the training set: for example, we calculated ECG feature quantiles using only training data. In total, the training set consisted of 4423 hospital admissions from 4423 distinct patients (1293 with LOS >14 days, 782 with mortality events). The testing set consisted of 491 patients (91 with LOS >14 days, 63 with mortality events).

5. Results

5.1. Topic interpretation

For many of the latent topics found by Wave-MixEHR, the most-associated ICD codes, prescriptions, lab abnormalities, and note keywords reflect a specific disease alongside common complications and comorbidities. For example, topic number 1 reflects diabetes and its many complications, particularly diabetic neuropathy. The most-associated ICD code with this topic was 25060 ("Diabetes with neurological manifestations"), with other associated ICD codes for diabetic ketoacidosis, diabetic retinopathy, diabetic kidney disease, foot ulcers, and cellulitis. The clinical notes keyword most associated with this topic was "foot," likely referring to diabetic foot ulcers, a common complication caused by diabetic neuropathy (Forbes and Cooper, 2013). Other associated note keywords included "sugars," "hyperglycemia," and insulin formulations "humalog," "lantus," and "glargine." Insulin did not appear in the associated prescriptions, possibly having been filtered out during preprocessing as too common to be informative. Prescriptions associated with this topic included gabapentin (used to treat painful diabetic neuropathy (Bennett and Simpson,

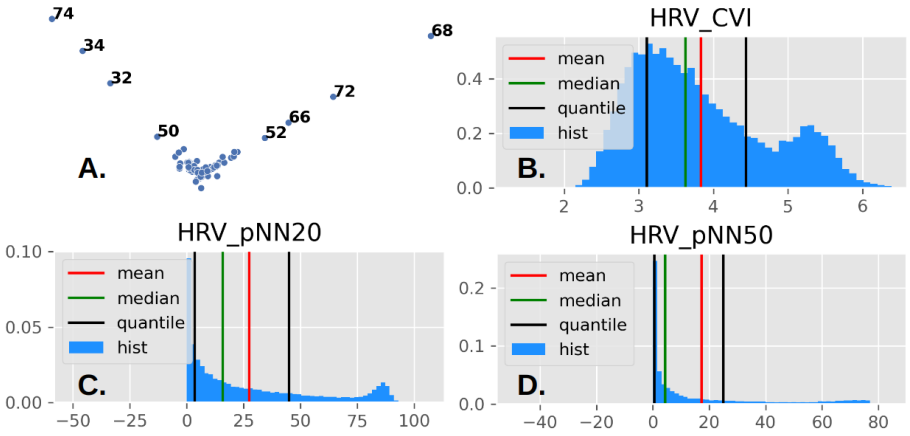

*Figure 2: Some seemingly ECG-driven outlier topics correspond with bimodally-distributed ECG features. Panel **A** is a PCA plot of Wave-MixEHR's 75 topics, based on each topic's level of association with each of the 246 ECG quantile features in our dataset. Outlier topics are labelled by number. Panels **B**, **C**, and **D** show the training set distributions of the top three ECG-derived features associated with topic number 34. For all three of these features, topic 34 was associated with values above the 75th percentile (indicated by the right-most "quantile" line), roughly corresponding to the secondary modes observed at high values of these features. Topic 32, an outlier positioned near topic 34 in the PCA plot, is also closely associated with high values of the features in panels **B** and **C**.*

2004)), metoclopramide (used to treat diabetic gastroparesis (Shakhatreh et al., 2019)), blood pressure-lowering drugs labetalol, valsartan, and carvedilol (hypertension can be both a comorbidity and a complication of diabetes (Lago et al., 2007)), and both ferrous sulfate and epoetin for treating anemia, another diabetes complication by way of diabetic kidney disease (Thomas, 2007). Appropriately, the abnormal laboratory test most associated with this topic was hemoglobin A1c. As another example, the note keywords most associated with topic number 9 are "cirrhosis," "portal," "ascites," "TIPS," and "varices." A recognizable story emerges: **cirrhosis** of the liver causes **portal** vein hypertension, which in turn causes **ascites** and esophageal **varices** (Maruyama and Yokosuka, 2012). Bleeding from esophageal varices is a common complication: one possible treatment is a **TIPS** procedure, in which the portal vein is connected with a nearby systemic vein to relieve portal hypertension (Rössle and Grandt, 2004). We found other topics that seemed to clearly represent pacemaker implantation, hip replacement surgery, atrial fibrillation, HIV, empyema, myocardial infarction, cancer metastasis, stroke, seizures, sepsis, head trauma, prosthetic heart valves, coronary bypass surgery, motor vehicle accidents, and many other concepts.

Interpretation of ECG feature associations was more challenging and yielded mixed results. Many of the features are correlated with each other due to their association with heart rate variability, which can be measured from ECG signals by a variety of methods, and low values of which have been associated with a range of negative health outcomes including cardiac mortality (Karim et al., 2011).

One approach we took to identify meaningful ECG topic associations was to first use features from non-ECG modalities to manually identify cardiac-related topics, before examining ECG features associated with those topics. Topic number 16, visualized using word clouds in panels **A**-**E** of figure 3, is associated with ICD codes and note keywords for

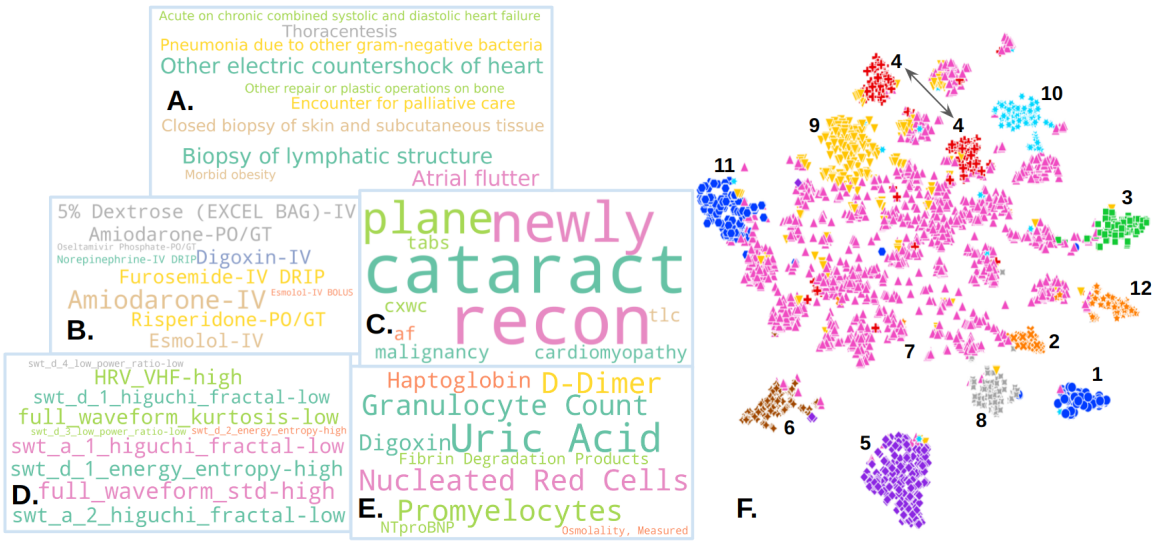

*Figure 3: Many latent topics represent specific diseases or closely related comorbidities, while clusters of patients in topic-space seem to represent higher-level concepts such as entire organ systems. Word clouds for topic number 16 reveal expected and unexpected associations with atrial fibrillation. Panels **A** (ICD9 codes), **B** (prescriptions), **C** (clinical note keywords), **D** (ECG features) and **E** (abnormal lab tests) show the top 10 most-associated features with this topic for each modality. Panel **F** shows patients clustered by their membership in each of Wave-MixEHR's 75 topics. K-means clustering assigned cluster labels (shown near each cluster in the plot) before t-SNE mapped each patient to one point on this two-dimensional plot.*

atrial fibrillation (AF), atrial flutter, cardioversion treatment for AF, and prescriptions for amiodarone (which treats AF) (Lip and Tse, 2007). This topic was also associated with a high standard deviation of the ECG waveform, likely corresponding to the large ECG oscillations caused by rapid atrial contractions during atrial flutter (Cosío, 2017). The second most-associated ECG feature with topic 16 is a measure of energy entropy in the stationary wavelet transform, which was used by a previous machine learning study to accurately detect AF (Asgari et al., 2015). Interestingly, the lab test result most closely associated with this topic is abnormal serum uric acid, which was recently found to be associated with increased risk of new-onset AF (Ding et al., 2023). Another unexpected finding is that "cataract" is the top clinical note keyword associated with this topic. A literature search yielded a large-scale retrospective cohort study showing that patients with cataracts are at higher risk of developing AF than those without, independently of age and various other risk factors (Hu and Lin, 2017). Several other cardiac-related topics also have ECG standard deviation and/or stationary wavelet transform energy entropy as some of their most closely-associated ECG features: topics 7 (with other features indicating intra-aortic balloon pump), 23 (cardiac tamponade), 45 (systolic heart failure) and 55 (arrhythmias and cardiac ablation).

Some topics had interpretable ECG feature associations despite not being directly related to cardiac conditions. Topic number 10 has compartment syndrome, rhabdomyolysis, and fasciectomy/fasciotomy as its most-associated ICD codes, and various formulations of prismasate as its top most-associated prescriptions. Prismasate is a renal dialysis fluid used to prevent and manage acute kidney failure, which is a common and serious compli-

cation of rhabdomyolysis and compartment syndrome (Galvagno Jr et al., 2013). These related conditions frequently result in release of intracellular potassium from damaged muscle tissue, causing hyperkalemia, which in turn causes heightened T-waves and widened QRS-complexes on ECG readings (Zimmerman and Shen, 2013). This could explain why topic 10 is also associated with a high power ratio for medium and low-frequency bands in ECG traces: enhanced T waves and widened QRS could be represented by lower-frequency information than a normal ECG.

Another relatively interpretable assocation is that between high SDNN and SDSD (measures of heart rate variability, (Shaffer and Ginsberg, 2017)) and topic 35, which has "cardiac arrest" and "cardiopulmonary rescuscitation" as its top two most-associated ICD codes, morphine as its top prescription, and top note keywords "arrest," "asystolic," "PEA" (pulseless electrical activity), "expired," and "CMO" (comfort measures only). It's logical that patients who have died would have (for example) a high standard deviation of the inter-beat interval on ECG (SDNN) due to potential periods of asystole in the recording.

Another approach we took to interpreting ECG feature associations latent topics was to conduct principle component analysis (PCA) on the 75 topics based entirely on their associations with various ECG features, as shown in figure 2A. We observed that, in contrast to the majority of the topics in our model, most of the outlier topics in figure 2 did not have obvious clinical interpretations as they were associated with sets of seemingly unrelated ICD codes and other features. Topic 34, however, appears to be associated with biliary tract obstruction, with top ICD codes for percutaneous biliary tract procedures, cholangitis, obstruction of bile duct, and cholangiograms, as well as note keywords "duct" and "bile" and abnormal lab tests for direct bilirubin (which can be caused by bile duct obstruction, Méndez-Sánchez et al. (2019)). Topic 34 is also associated with high values of multiple bimodally distributed ECG features, corresponding to outlying modes (see figure 2). HRV_CVI, the top ECG feature for topic 34, refers to the cardiac vagal index, a feature derived from heart rate variability waveforms that is correlated with parasympathetic nervous system activity (Toichi et al., 1997). In the setting of bile duct obstruction and liver injury (which can result from bile duct obstruction), experimental studies in rodents have highlighted the important role of parasympathetic innervation from the vagus nerve in reducing liver inflammation, facilitating hepatic repair following liver injury, and both decreasing apoptosis and increasing proliferation of cholangiocytes, the epithelial cells that line the bile ducts, following injury to the biliary tree (Miller et al., 2021). Mice with liver inflammation artificially induced by Fas antibody injection have significantly increased mortality when the vagus nerve branch carrying parasympathetic signals to the liver is severed (Hiramoto et al., 2008). We speculate that patients with bile duct obstruction may have increased parasympathetic tone as a protective response against hepatic injury, resulting in Wave-MixEHR's association between bile duct obstruction and the cardiac vagal index.

Clustering patients by their associations with each topic (panel **F** of figure 3) reveals higher-level associations that in some cases seem to coalesce around specific organ systems. For example, cluster 8 seems to be about various injuries to the brain. This cluster is associated with topic 26, which is associated with icd codes for accidental head trauma, note keywords "subdural," "subarachnoid," "intraparenchymal" (for different types of brain hemorrhage), and prescriptions for likely seizure prophylaxis (levetiracetam, phenytoin, fosphenytoin). This same cluster of patients is also associated with topic 61 (primary and

metastatic brain tumors), topic number 25 (motor vehicle accidents, with an interesting association to abnormal blood alcohol lab tests), and topic 50 (ischemic stroke). Similarly, cluster 5 is associated with topic 70 (coronary bypass and heart valve replacements), topic 45 (heart failture and prosthetic heart valves), topic 21 (myocardial infarction with and without ST-segment elevation), and topics 72 and 74, which do not have obvious clinical interpretations but appear as ECG-related outlier topics in panel A of figure 2.

5.2. Predicting mortality and length-of-stay

The performance on both binary prediction tasks of survival and length-of-stay are shown in 1. From the results, we see that incorporating ecg features both into wave-mixehr and outside of it consistently improves performance on both tasks. We show in figure 4 that introduction of ecg features through intermediate fusion alters the relative importance of topic features according to the XGBoost model. Interestingly, heart rate variablity is consistently a driving factor in the final model predictions. Among the outlier features identified in figure 2 HRV_CVI (cardiac vagal index, a measure of parasympathetic tone) ranks #1, HRV_pNN50 ranks #7, and HRV_pNN20 ranks #11 on the the LOS model trained on ECG features alone.

Task	Modality	Setting	auc	accuracy	f1
SRV	ecg feats	ecg	0.709	0.840	0.337
SRV	topics	topics(labs)	0.562	0.632	0.230
SRV	topics	topics(labs,notes)	0.906	0.756	0.478
SRV	topics	topics(labs,notes,ecg)	0.933	0.872	**0.618**
SRV	intermed fusion	ecg feats + topics(labs)	0.528	0.872	0.00
SRV	intermed fusion	ecg feats + topics(labs,notes)	0.907	0.718	0.353
SRV	intermed fusion	ecg feats + topics(labs,notes,ecg)	**0.944**	0.748	0.374
SRV	late fusion	ecg feats + topics(labs)	0.548	**0.882**	0.174
SRV	late fusion	ecg feats + topics(labs,notes)	0.598	0.872	0.061
SRV	late fusion	ecg feats + topics(labs,notes,ecg)	0.619	0.876	0.063
Task	Modality	Setting	auc	accuracy	f1
LOS	ecg feats	topics(ecg)	0.566	0.866	0.042
LOS	topics	topics(labs)	0.475	0.244	0.231
LOS	topics	topics(labs, notes)	0.814	0.315	0.291
LOS	topics	topics(labs,notes,ecg)	0.760	0.199	0.262
LOS	intermed fusion	ecg feats + topics(labs)	0.528	0.872	0.000
LOS	intermed fusion	ecg feats + topics(labs,notes)	0.907	0.718	0.353
LOS	intermed fusion	ecg feats + topics(labs,notes,ecg)	**0.941**	**0.729**	0.301
LOS	late fusion	ecg feats + topics(labs)	0.578	0.626	0.369
LOS	late fusion	ecg feats + topics(labs,notes)	0.553	0.657	0.352
LOS	late fusion	ecg feats + topics(labs,notes,ecg)	0.576	0.622	**0.419**

Table 1: Performance results on the held-out testing set for all run settings. Task indicates whether the binary prediction task was survival (SRV) or length-of-stay > 7 days. Modality indicates whether the input to the model was ecg features (ecg feats), topic models from mixehr (topics), a fusion of topics and ecg features (intermed fusion), or an averaged prediction of two separate models (late fusion). Setting indicates the data types used for a given modality, with "+" indicating concatenated features. The parenthetical text indicates the modalities provided to the topic model. The best performance for each metric across each task is shown in bold. Note that for brevity, "labs" is used to represent both lab results and prescription information.

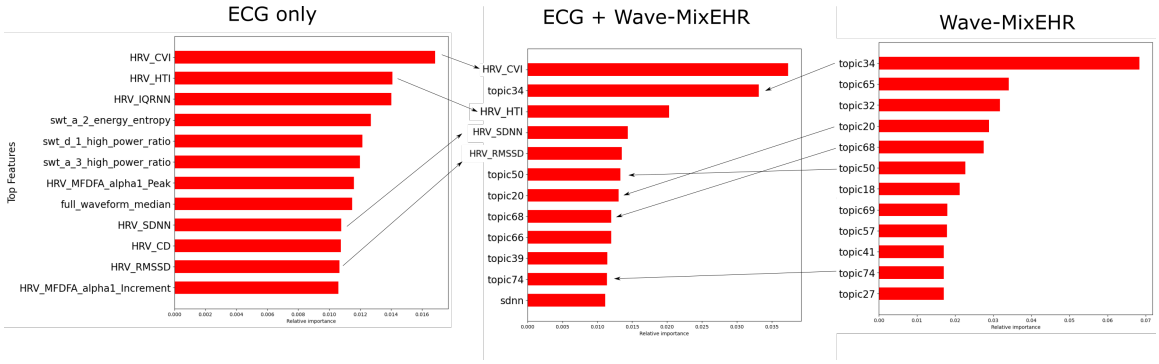

Figure 4: Top 12 highest importance features for XGBoost features of mortality prediction on ECG features alone, intermediate fusion of ECG features with Wave-MixEHR topics, and Wave-MixEHR topics alone. Feature importance was computed according to average entropy reduction between a parent node's entropy and the weighted sum of the child nodes' entropy after splitting on the feature. Arrows indicate changes in feature importance from the Wave-MixEHR model and ECG-only models, and the combined topic+ecg feature model after intermediate fusion.

6. Discussion

We trained a generative latent topic model on EHR data from the MIMIC-III Waveform Database Matched Subset, incorporating continuous waveform recordings for the first time in the form of ECG traces converted to categorical representations. We then used ECG information and the topic distribution for each patient to predict mortality and length-of-stay. Most of the latent topics are interpretable, typically associated with multimodal features of a specific disease as well as common comorbidities and complications. Our findings support the conclusions of relatively recent cohort studies with tens to hundreds of thousands of patients, demonstrating that the risk of new-onset atrial fibrillation (AF) is associated with cataracts (Hu and Lin, 2017) and hyperuricemia (Ding et al., 2023), independently of age and a variety of other risk factors and potential confounders. For specific examples of both cardiac (e.g., AF) and non-cardiac (e.g., rhabdomyolysis) topics, we also found clinically meaningful associations with ECG features related to heart rate variability and frequency power bands. The fact that Wave-MixEHR identified straightforward ECG feature associations, such as AF's association with increased waveform standard deviation and cardiac arrest's association with various anomalous readings, suggests that some of the more unexpected associations may warrant further investigation. In particular, we found an intruiging association between bile duct obstruction and increased parasympathetic tone as measured by the ECG-derived cardiac vagal index (topic 34). A wide array of rodent studies have highlighted the critical role of parasympathetic innervation from the vagus nerve in facilitating liver and biliary tract repair following injury (e.g., due to bile duct obstruction, Miller et al. (2021)). Our findings provide preliminary correlational evidence for the possible role of parasympathetic signalling following bile duct obstruction in humans, which to the best of our knowledge has not previously been studied directly. The fact that this same topic 34 and the ECG cardiac vagal index associated with it were the top-two most important features by far in predicting patient mortality (figure 4) lends urgency to this line of inquiry.

Our evaluation of supervised models on different modalities and fusion methods revealed that the Wave-MixEHR model trained on lab results, prescription information, clinical notes, and ecg features consistently outperformed models without context from all modalities. In particular, ECG features alone performs relatively poorly on the mortality (auc=0.709, f1=0.337) and LOS tasks (auc=0.566, f1=0.042), yet the addition of ecg features to MixEHR in both early and intermediate fusion substantially improves performance across both tasks. These results suggest that ECG alone may not contain sufficient information to predict patient outcome, but it provides valuable context when analyzed in conjunction with other clinical modalities. Specifically, the inclusion of ECG features into MixEHR increased auc from 0.906 to 0.933, F1 from 0.478 to 0.618 on survival prediction, though inclusion of ecg features moderately decreased performance in LOS prediction from 0.814 auc to 0.760 in early fusion approaches. However, intermediate fusion of ECG features alongside Wave-MixEHR topics consistently leads to the highest-performing model across both tasks in terms of AUC.

Our comparison of early, intermediate, and late fusion provides further insight into the nature of the topic model. Intuitively, introducing ECG features earlier into the model allows for more intricate feature interactions to be learned: however, introducing the new ECG modality directly into MixEHR also makes the learning problem more prone to overfitting, as we vastly increase dimensionality without guarantee that the ECG features confer clinically relevant information. Furthermore, clinically-relevant details for our supervised tasks from ECG features may not be reflected by topic assignments due to the unsupervised nature of topic models. In contrast, the late fusion model which simply ensembles the predictions between MixEHR and ECG-alone circumvents the need to learn the relationship between ECG features and other modalities, ensuring the ECG features are used for the final prediction. The intermediate fusion model serves as a compromise between these two approaches, preserving the signal from the ECG features at the cost of higher dimensional inputs for XGBoost. As a whole, we observe from 1 that late fusion approaches consistently *underperform* early- and intermediate- fusion approaches, indicating that the interaction between ECG and MixEHR contains clinically-relevant information for patient outcome prediction. Furthermore, we see that for mortality prediction, early fusion of all modalities leads to comparable performance with intermediate fusion of Wave-MixEHR with ecg features, suggesting that Wave-MixEHR is able to encode the majority of clinical information contained within the ecg features at a much smaller dimensionality (75 features vs 225 features).

Our study has several important limitations. In our encoding of ECG features, we incorporated one 10-second waveform in each hour of a patient's admission: while this approach may capture relatively stable ECG waveform characteristics, it may miss transient events such as certain arrhythmias. Interpretation of clinical feature associations with latent topics cannot establish causal relationships among features, although some known causal relationships such as that between diabetic neuropathy and foot ulcers were identified. Future work could combine multimodal topic modeling with causal modeling: for example, by training deep twin networks to answer counterfactual questions about categorical associations identified by MixEHR models (Vlontzos et al., 2023).

Acknowledgments and Member Contributions

We would like to acknowledge our clinical and research mentors Yue Li, Ph.D., and Manolis Kellis, Ph.D., for their help and support in guiding the direction of our project and our research questions.

Payal Chandak performed ECG signal processing, breaking up the signals into 12 hour trajectories, exploring the data distribution and identifying the key lead II from which we focused our analysis, extracted 150 handcrafted features from each trajectory, and organized the MIMIC III dataset into train/test splits.

Daniel Shao performed supervised prediction of ECG features from Payal and MixEHR topics from Morgan for LOS and mortality prediction. He implemented early fusion, intermediate fusion, and late fusion with various permutations of topic model-ecg feature pairs to obtain greater insights into the nature of how ECG features interact with other clinical data. He generated feature importance plots from XGBoost to further interpret the topics and relationships between features.

Joanna Shen implemented a pipeline to identify the top 10 lab tests associated with each of 75 topics and generate identifiers for lab tests under each topic. Combining these identifiers with the ones Morgan obtained for ICD-9 codes, prescriptions, lab tests, and clinical notes for each topic, she used the WordCloud library in Python to generate word clouds which help improve the interpretability of this project. The word clouds were generated based on the frequencies of each identifier among the top 20 phenotypes. She helped wrote the introduction and related work section of this report.

Morgan Talbot implemented a pipeline to pre-process non-ECG data from MIMIC-III (ICD9 codes, prescriptions, lab tests, and clinical notes) in order to train the Wave-MixEHR model. Building on Payal's ECG feature development and extraction work, he developed a strategy to convert ECG features to binarized quantile memberships by visualizing the feature distributions. Building on Joanna's visualizations of latent topic associations with phenotypes such as ICD codes, ECG features, etc., he applied his knowledge from (the first half of) medical school to interpret multimodal latent topic associations, including detailed analyses of topics representing diabetes, cirrhosis, atrial fibrillation, and bile duct obstruction, with references to studies that substantiate the multimodal associations of these topics. He also performed k-means clustering and t-SNE visualization of patients in topic-space to identify clusters associated with (as examples) brain injury and heart disease, and generated a PCA plot to identify latent topics that were outliers in terms of their ECG feature associations.

7. Appendix

7.1. Feature Description

- **Time-Frequency Analysis.** In the time domain, we can analyse quantities such as the minimum, maximum, median, mean, variance, skew, kurotsis, etc. To transfer the ECG into frequency domain, we apply a stationary wavelet transform (SWT). The SWT provides a shift-invariant representation of signals by decomposing them into wavelet coefficients at multiple scales and locations (Nason and Silverman, 1995). When applied to ECG data, SWT can effectively isolate and enhance signal com-

ponents associated with specific cardiac events, thus improving the detection and analysis of clinically relevant features (Goodfellow et al., 2018). Furthermore, the shift-invariant property of SWT guarantees robustness against variations in the ECG signal's time alignment, which enables reliable extraction of diagnostic information from the data (Nason and Silverman, 1995).

  - **Power ratios.** Using the power ratio of the SWT as a feature can help identify specific cardiac events by capturing the relative energy distribution across different frequency bands (**?**). By comparing the ratios of frequency components, events such as atrial fibrillation, ventricular arrhythmias, and myocardial ischemia can be effectively identified, as they exhibit distinct power distribution patterns across various frequency scales (Asgari et al., 2015).
  - **Energy entropy.** Energy entropy effectively represents the frequency distribution and randomness inherent in ECG signals (Acharya et al., 2005; Asgari et al., 2015). A high entropy value signifies increased complexity or randomness, whereas a low value implies a more organized or regular pattern. This set of features can not only help the model distinguish noise from ECG, but also vary with different types of arrhythmias (Acharya et al., 2005; Asgari et al., 2015).
  - **Higuchi fractal.** The Higuchi fractal dimension captures information related to the ECG's frequency complexity and enables the differentiation of distinct patterns that may be correlated with various morbidities (Acharya et al., 2005).

- **Clinical Knowledge.** The ECG waveform exhibits a medically significant pattern that repeats with each heartbeat. This pattern comprises the P wave, representing atrial depolarization, followed by the QRS complex, which corresponds to ventricular depolarization, and concludes with the T wave, associated with repolarization (Camm et al., 2009). Clinicians frequently diagnose cardiovascular diseases based on the relative relationships between these attributes within a single heartbeat (e.g., ST elevation, interbeat intervals) and across multiple beats (e.g., RR intervals) (Camm et al., 2009). We build a set of clinical features with assistance from PySiology (Esposito et al., 2020) and NeuroKit (Makowski et al., 2021).

  - **Beats.** We identified various points of interest for each beat, specifically the start of the P wave, R peaks, and start of the T waves. Subsequently, we estimated parameters including beats per minute, mean interbeat intervals, standard deviation of interbeat intervals, and others.
  - **R-R intervals.** By studying the time difference between consecutive R-R intervals, we estimated the number of intervals that differ from one another by more than 50 ms and 20 ms, as well as the root mean square of successive differences, among other features.
  - **Heart rate variability.** Heart rate variability refers to the changes in time intervals between consecutive heartbeats, reflecting the balance between the sympathetic and parasympathetic nervous systems that regulate the heart's function (van Ravenswaaij-Arts et al., 1993; Acharya et al., 2005). HRV can provide insight into how well the heart adapts to various physiological and environmental

factors (van Ravenswaaij-Arts et al., 1993). Given an ECG, we can construct a new HRV waveform and estimate various features from time and frequency domains as we did previously for the ECG waveform (Acharya et al., 2005).

References

Rajendra Acharya, P Subbanna Bhat, N Kannathal, Ashok Rao, and Choo Min Lim. Analysis of cardiac health using fractal dimension and wavelet transformation. *Itbm-Rbm*, 26(2):133–139, 2005.

Yuri Ahuja, Yuesong Zou, Aman Verma, David Buckeridge, and Yue Li. Mixehr-guided: A guided multi-modal topic modeling approach for large-scale automatic phenotyping using the electronic health record. *Journal of Biomedical Informatics*, 134:104190, 2022. doi: 10.1016/j.jbi.2022.104190.

Shadnaz Asgari, Alireza Mehrnia, and Maryam Moussavi. Automatic detection of atrial fibrillation using stationary wavelet transform and support vector machine. *Computers in biology and medicine*, 60:132–142, 2015.

Michael I Bennett and Karen H Simpson. Gabapentin in the treatment of neuropathic pain. *Palliative Medicine*, 18(1):5–11, 2004.

A John Camm, Thomas F Lüscher, and Patrick W Serruys. *The ESC textbook of cardiovascular medicine*. OXFORD university press, 2009.

Francisco G Cosío. Atrial flutter, typical and atypical: a review. *Arrhythmia & electrophysiology review*, 6(2):55, 2017.

Mozhu Ding, Ngoc Nguyen Viet, Bruna Gigante, Viktor Lind, Niklas Hammar, and Karin Modig. Elevated uric acid is associated with new-onset atrial fibrillation: Results from the swedish amoris cohort. *Journal of the American Heart Association*, page e027089, 2023.

Anna Esposito, Marcos Faundez-Zanuy, Francesco Carlo Morabito, and Eros Pasero, editors. *Neural Approaches to Dynamics of Signal Exchanges*. Springer Singapore, 2020. doi: 10.1007/978-981-13-8950-4. URL https://doi.org/10.1007/978-981-13-8950-4.

Josephine M Forbes and Mark E Cooper. Mechanisms of diabetic complications. *Physiological reviews*, 93(1):137–188, 2013.

Samuel M Galvagno Jr, Caron M Hong, Matthew E Lissauer, Andrew K Baker, Sarah B Murthi, Daniel L Herr, and Deborah M Stein. Practical considerations for the dosing and adjustment of continuous renal replacement therapy in the intensive care unit. *Journal of critical care*, 28(6):1019–1026, 2013.

A. Goldberger, L. A. Amaral, L. Glass, Jeffrey M. Hausdorff, P. Ivanov, R. Mark, J. Mietus, G. Moody, C. Peng, and H. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101 23:E215–20, 2000.

Sebastian D Goodfellow, Andrew Goodwin, Robert Greer, Peter C Laussen, Mjaye Mazwi, and Danny Eytan. Atrial fibrillation classification using step-by-step machine learning. *Biomedical Physics &amp Engineering Express*, 4(4):045005, May 2018. doi: 10.1088/2057-1976/aabef4. URL https://doi.org/10.1088/2057-1976/aabef4.

Scott Halgrim, Fei Xia, Imre Solti, Eithon Cadag, and Özlem Uzuner. A cascade of classifiers for extracting medication information from discharge summaries. *Journal of Biomedical Semantics*, 2(Suppl 3), Jul 2011. doi: 10.1186/2041-1480-2-s3-s2.

Tetsuya Hiramoto, Yoichi Chida, Junko Sonoda, Kazufumi Yoshihara, Nobuyuki Sudo, and Chiharu Kubo. The hepatic vagus nerve attenuates fas-induced apoptosis in the mouse liver via α7 nicotinic acetylcholine receptor. *Gastroenterology*, 134(7):2122–2131, 2008.

Wei-Syun Hu and Cheng-Li Lin. Association between cataract and risk of incident atrial fibrillation: a nationwide population-based retrospective cohort study. In *Mayo Clinic Proceedings*, volume 92, pages 370–375. Elsevier, 2017.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Jelena Jovanovic, Ebrahim Bagheri, John Cuzzola, Dragan Gasevic, Zoran Jeremic, and Reza Bashash. Automated semantic tagging of textual content. *IT Professional*, 16(6): 38–46, 2014. doi: 10.1109/MITP.2014.85.

Nasim Karim, Jahan Ara Hasan, and Syed Sanowar Ali. Heart rate variability-a review. *Journal of Basic & Applied Sciences*, 7(1), 2011.

Rodrigo M Lago, Premranjan P Singh, and Richard W Nesto. Diabetes and hypertension. *Nature clinical practice Endocrinology & metabolism*, 3(10):667–667, 2007.

Yue Li and Manolis Kellis. A latent topic model for mining heterogenous non-randomly missing electronic health records data, 2018.

Yue Li, Pratheeksha Nair, Xing Han Lu, Zhi Wen, Yuening Wang, Amir Ardalan Kalantari Dehaghi, Yan Miao, Weiqi Liu, Tamas Ordog, Joanna M Biernacka, et al. Inferring multimodal latent topics from electronic health records. *Nature communications*, 11(1): 2536, 2020.

Gregory YH Lip and Hung-Fat Tse. Management of atrial fibrillation. *The Lancet*, 370 (9587):604–618, 2007.

Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.

Chongyuan Luo, Christopher L. Keown, Laurie Kurihara, Jingtian Zhou, Yupeng He, Junhao Li, Rosa Castanon, Jacinta Lucero, Joseph R. Nery, Justin P. Sandoval, and et al. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science*, 357(6351):600–604, 2017. doi: 10.1126/science.aan3351.

Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. NeuroKit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696, feb 2021. doi: 10.3758/s13428-020-01516-y. URL https://doi.org/10.3758%2Fs13428-020-01516-y.

Hitoshi Maruyama and Osamu Yokosuka. Pathophysiology of portal hypertension and esophageal varices. *International journal of hepatology*, 2012, 2012.

Finale Doshi-Velez Nicole Brimmer Rohit Joshi Anna Rumshisky Marzyeh Ghassemi, Tristan Naumann and Peter Szolovits. Unfolding physiological state: Mortality modelling in intensive care units.

Nahum Méndez-Sánchez, Xingshun Qi, Libor Vitek, and Marco Arrese. Evaluating an outpatient with an elevated bilirubin. *Official journal of the American College of Gastroenterology— ACG*, 114(8):1185–1188, 2019.

Bess M Miller, Isaac M Oderberg, and Wolfram Goessling. Hepatic nervous system in development, regeneration, and disease. *Hepatology*, 74(6):3513–3522, 2021.

Benjamin Moody, George Moody, Mauricio Villarroel, Gari Clifford, and Ikaro Silva. Mimic-iii waveform database matched subset, 2020. URL https://physionet.org/content/mimic3wdb-matched/1.0/.

Guy P Nason and Bernard W Silverman. The stationary wavelet transform and some statistical applications. *Wavelets and statistics*, pages 281–299, 1995.

Martin Rössle and Daniel Grandt. Tips: an update. *Best Practice & Research Clinical Gastroenterology*, 18(1):99–123, 2004.

Fred Shaffer and Jay P Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in public health*, page 258, 2017.

Mohammed Shakhatreh, Asad Jehangir, Zubair Malik, and Henry P Parkman. Metoclopramide for the treatment of diabetic gastroparesis. *Expert Review of Gastroenterology & Hepatology*, 13(8):711–721, 2019.

Ziyang Song, Xavier Sumba Toral, Yixin Xu, Aihua Liu, Liming Guo, Guido Powell, Aman Verma, David Buckeridge, Ariane Marelli, and Yue Li. Supervised multi-specialist topic model with applications on large-scale electronic health record data. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384506. doi: 10.1145/3459930.3469543. URL https://doi.org/10.1145/3459930.3469543.

Li Sun, Yanping Lu, Kaitao Yang, and Shaozi Li. Ecg analysis using multiple instance learning for myocardial infarction detection. *IEEE Transactions on Biomedical Engineering*, 59(12):3348–3356, 2012. doi: 10.1109/TBME.2012.2213597.

Merlin C Thomas. Anemia in diabetes: marker or mediator of microvascular disease? *Nature clinical practice Nephrology*, 3(1):20–30, 2007.

Motomi Toichi, Takeshi Sugiura, Toshiya Murai, and Akira Sengoku. A new method of assessing cardiac autonomic function and its comparison with spectral analysis and coefficient of variation of r–r interval. *Journal of the autonomic nervous system*, 62(1-2): 79–84, 1997.

Alexander Van Esbroeck, Chih-Chun Chia, and Zeeshan Syed. Heart rate topic models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 26(1):1635–1641, 2021. doi: 10.1609/aaai.v26i1.8337.

Conny MA van Ravenswaaij-Arts, Louis AA Kollee, Jeroen CW Hopman, Gerard BA Stoelinga, and Herman P van Geijn. Heart rate variability. *Annals of internal medicine*, 118(6):436–447, 1993.

Athanasios Vlontzos, Bernhard Kainz, and Ciarán M Gilligan-Lee. Estimating categorical counterfactuals via deep twin networks. *Nature Machine Intelligence*, pages 1–10, 2023.

Janice L Zimmerman and Michael C Shen. Rhabdomyolysis. *Chest*, 144(3):1058–1065, 2013.