Phenotyping with Partially Labeled, Partially Observed Data

Victor A. Rodriguez

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2023

Abstract

Phenotyping with Partially Labeled, Partially Observed Data

Victor A. Rodriguez

Identifying a group of individuals that share a common set of characteristics is a conceptually simple task, which is often difficult in practice. Such *phenotyping* problems emerge in various settings, including the analysis of clinical data. In this setting, phenotyping is often stymied by persistent data quality issues. These include a lack of reliable labels to indicate the presence of absence of characteristics of interest, and significant missingness in observed variables. This dissertation introduces methods for learning phenotypes when the data contain missing values (partially observed) and labels are scarce (partially labeled). Aim 1 utilizes an unsupervised probabilistic graphical model to learn phenotypes from partially observed data. Aim 2 introduces a related semi-supervised probabilistic graphical model for learning phenotypes from partially labeled clinical data. Finally, Aim 3 describes a method for training deep generative models when the training data contain missing values. The algorithm is then applied in a semi-supervised setting where it accounts for partially labeled data as well.

Table of Contents

Acknowledgements

This dissertation would not have been possible without the help and support contributed by many individuals. Among them, first I would like to express my gratitude to my advisor, Dr. Adler Perotte. I learned so much during my time as your student. Thank you for sharing your knowledge, your interest, and, perhaps most importantly, your patience.

I would like to thank the members of my dissertation committee. Dr. George Hripcsak and Dr. Nicholas Tatonetti, thank you for you insights and guidance throughout my graduate years and for helping to shepherd me through my dissertation work. Dr. Marissa Burgermaster, thank you for you constant support, mentorship, and the occassional pep talks. Dr. Rahul Krishnan, thank you for signing on to this committee and bringing your expertise and experience to bear.

Thank you also to all my colleagues at Columbia's Department of Biomedical Informatics. Dr Inigo Urteaga, thank you for sharing what you know, and for helping me think through my ideas. Dr. Amelia Averitt, thank you for always being willing to field my questions on courses, research, and navigating through graduate life. Shreyas and Mert, thank you for your interest in my work, and for affording me the opportunities to get involved in yours.

Finally, I would like to thank my friends and family. To my wife, Dr. Sarah McKetta, thank you for your untiring support. I really, truly, could not have gotten here without you. And everyone else, thank you for putting up with ramblings these past many years and for being patient with me as I made my way through.

Dedication

To my mother and father who started me moving and my wife who kept me going.

Chapter 1

Introduction

Problem Statement

The adoption of electronic health records has permitted the large scale collection and analysis of observational clinical datasets. A wide variety of research has since emerged, driven by the richness and scale these data possess. Common to many of these works is a persistent bottleneck encountered early data processing: the phenotyping problem. In simple terms, the phenotyping problem has to do with identifying the patients one, as a researcher, is interested in studying. Patients in such a cohort may share a common disease, or perhaps a common clinical intervention, whatever the researcher is interested in. At first glance, phenotyping seems like no problem at all; just query the data for whatever clinical feature is of interest and isolate all the patients that have it. However, this perspective does not stand up to the realities of clinical data.

In essence, phenotyping is hard because of what clinical data are for; they are meant for documentation, not research. Thus, they lack certain features that would otherwise make phenotyping trivial. Most prominently, clinical data generally do not contain reliable, gold standard labels for the things researchers are most often interested in (e.g. disease diagnoses). These data also routinely contain significant amounts of missingness. Since clinicians record only a subset of clinical variables for a given patient at a given time, most variables go unobserved.

Given these obstacles, many researchers resort to defining phenotyping algorithms using hand written rules. However, because phenotyping is essentially a classification task, there

is much interest in using supervised models to learn phenotypes directly from data. Unfortunately, supervised models commonly assume the labels and data are fully observed which complicates their application to clinical problems, including phenotyping.

In this work we propose several methods for learning phenotypes from clinical data. We develop models and algorithms for training them, which assume the data and labels are only partially observed. By recognizing the realities of clinical data from the outset, we aim to deliever effective phenotyping methods to the biomedical informatics community to alleviate the bottleneck represented by the phenotyping problem.

Purpose of the Study

This work proposes methods for learning phenotypes from clinical data. Specifically, it focuses on phenotyping when the data are partially observed and partially labeled — a typical scenario when working with observational clinical data.

Aim 1 utilizes an unsupervised probabilistic graphical model to learn phenotypes from partially observed, heterogeneous clinical data. Aim 2 builds on Aim 1 by proposing a related semi-supervised probabilistic graphical model for learning interpretable phenotypes from partially labeled clinical data. Finally, Aim 3 describes a method for training deep generative models when the training data contain missing values. The algorithm is then applied in a semi-supervised setting where training occurs on partially labeled data as well.

Research Questions and Hypotheses

Aim 1. Implement a method for learning phenotypes from partially observed data.

Aim 1A. ***Utilize the Multi-Channel Mixed Membership Model (MC3M) to learn psychosocial-behavioral phenotypes from partially observed survey data***

Research Question	*Can MC3M identify subgroups of individuals that exhibit similar psycho-social traits?*
Hypothesis	Individuals with similar traits will have similar subsets of active phenotypes; moreover, any individual's observations will be explained by a relatively small number of active phenotypes.

Aim 1B. ***Identify MC3M phenotypes associated with a health outcome.***

Research Question	*Can MC3M phenotypes be used to predict the presence of a targeted health outcome?*
Hypothesis	When used as covariates in a predictive model, a subset of MC3M phenotypes will be predictive of elevated weight status (EWS: BMI $> 25\text{kg/mg}^2$)

Aim 2. Develop a method for learning phenotypes from partially labeled clinical data.

Aim 2A. ***Derive and implement the Semi-Supervised Mixed Membership Model (SS3M) — a probabilistic graphical model for learning interpretable, disease-specific phenotypes from partially labeled, multi-modal clinical data.***

Research Question	*Can SS3M recover ground-truth phenotypes from partially labeled data?*
Hypothesis	Given simulated data containing a subset of labeled examples, SS3M will recover the structure of ground-truth phenotypes. Moreover, the identity of each phenotype will correctly match the corresponding label.

Aim 2B. ***Utilize SS3M to learn disease-specific phenotypes from partially labeled observational clinical data.***

Research Question	*Do SS3M phenotypes capture the clinical characteristics of the diseases specified by the labels provided?*
Hypothesis	Clinical experts will find SS3M phenotypes better represent specific diseases when compared to phenotypes learned using an unsupervised baseline. In addition, experts will find the content of SS3M phenotypes is representative of the diseases specified by their labels.
Research Question	*Is SS3M an effective model for determining which patients have a specific disease?*
Hypothesis	Relative to supervised baselines, SS3M will demonstrate competitive or superior performance on disease label prediction.

Aim 3. Develop a method for learning phenotypes from partially observed, partially labeled data.

Aim 3A. ***Derive and implement a Monte Carlo Expectation-Maximization (MCEM) algorithm for training Variational Autoencoders (VAEs) on partially observed data.***

Research Question	*Is a VAE trained with MCEM an effective model for performing missing value imputation?*
Hypothesis	Using MCEM, a VAE trained on data with simulated missingness will produce higher quality imputations than those generated by baseline imputation algorithms.

Aim 3B. ***Develop a semi-supervised VAE which can be trained on partially observed data using MCEM; use it for disease phenotyping on clinical data with inherent missingness.***

Research Question	*Does a VAE trained on partially labeled, partially observed data yield an effective disease phenotyping model?*
Hypothesis	Relative to baselines comprising imputation followed by supervised learning, an MCEM trained semi-supervised VAE will yield superior or competitive performance.

Experimental Design Associated with Hypotheses

Aim 1. Implement a method for learning phenotypes from partially observed data.

Aim 1A. ***Utilize the Multi-Channel Mixed Membership Model (MC3M) to learn psychosocial-behavioral phenotypes from partially observed survey data*** We utilize an unsupervised phenotyping algorithm, MC3M, to learn phenotypes from partially observed survey data. We detail the structure of MC3M's probabilistic graphical and an inference algorithm which uses Gibbs sampling to learn the structure of latent phenotypes. We then train MC3M on tokenized survey data and interrogate the resultant phenotypes for their ability to identify meaningful subgroups of individuals within the survey sample population.

Aim 1B. ***Identify MC3M phenotypes associated with a health outcome.*** MC3M summarizes an individual's observations using a person-phenotype distribution — a probabilty vector representing the degree to which each phenotype explains the individual's data. Using the person-phenotype distributions as covariates, we train logistic regression models to predict a binary indicator for elevated weight status (EWS). We then use the regresssion coefficients to identify a subset of phenotypes showing significant assocations to EWS.

Aim 2. Develop a method for learning phenotypes from partially labeled clinical data.

Aim 2A. ***Derive and implement the Semi-Supervised Mixed Membership Model (SS3M) — a probabilistic graphical model for learning interpretable, disease-specific phenotypes from partially labeled, multi-modal clinical data.*** Disease phenotyping can often be reduced to classification. However, training supervised phenotyping models with electronic health records data is challenging due to the lack of gold standard la-

bels. Though fully labeling a clinical dataset is generally infeasible, obtaining a small amount of high-quality labels may be possible. Here we develop Semi-Supervised Mixed Membership Models (SS3M) — a family of semi-supervised models for learning phenotypes from partially labeled data. We build SS3M by incorporating a novel semi-supervision mechanism into an otherwise fully unsupervised model previously proposed for learning interpretable phenotypes from heterogenous clinical data. We derive and implement a Markov Chain Monte Carlo sampler to perform posterior inference on the model's latent variables. We then evaluate SS3M in simulation. First we simulate data from SS3M's generative model parameterized with a set of ground truth phenotypes and their corresponding labels. We then fit these data with a randomly initialized SS3M model and check to see if the model can recover the ground truth phenotypes in both structure and identity.

Aim 2B. ***Utilize SS3M to learn disease-specific phenotypes from partially labeled observational clinical data.*** We train SS3M on partially labled clinical data extracted from the MIMIC-III critical care database. We also train a closely related, fully unsupservised baseline on the same data, but without the labels. We recruit two clinical experts to evaluate the quality and content of phenotypes learned by both models. In addition, we evaluate SS3M's predictive performance relative to common supervised baselines, and explore how the performance is impacted by the amount of labeled data made available during training.

Aim 3. Develop a method for learning phenotypes from partially observed, partially labeled data.

Aim 3A. ***Derive and implement a Monte Carlo Expectation-Maximization (MCEM) algorithm for training Variational Autoencoders (VAEs) on partially observed data.*** Clinical data commonly contain missing values. Meanwhile, popular deep generative modeling frameworks like Variational Autoencoders (VAEs), often assume the training data are fully observed. Thus, the partially observed nature of clinical data complicates the

use of VAEs in clinical modeling tasks. In this work, we develop an algorithm for training VAEs on partially observed data. We derive a Monte Carlo Expectation-Maximization (MCEM) algorithm for VAEs, which (double) lower bounds the marginal log-likelihood of the observed data. VAEs trained using this algorithm are evaluated on a suite of missing value imputation tasks where imputation quality is evaluated by 1) visual inspection of imputed samples and 2) using two-sample Kolmogorov–Smirnov tests to measure the similarity between samples of missing values generated by the model and samples from the true conditional distributions $p(\boldsymbol{x}_m|\boldsymbol{x}_o)$, where $\boldsymbol{x}_o$ and $\boldsymbol{x}_m$ refer to observed and missing variables, respectively. VAE-based deep generative imputers serve as baselines for comparison.

Aim 3B. ***Develop a semi-supervised VAE which can be trained on partially observed data using MCEM; use it for disease phenotyping on clinical data with inherent missingness.*** Missingness in clinical datasets can occur in both the features and the labels. To train disease phenotyping models in this setting, we combine MCEM with semi-supervision. We show how to incorporate the training objective for a semi-supervised VAE into our MCEM algorithm. Using the Women in Data Science Datathon 2020 critical care dataset, we train the model to predict partially observed disease labels using clinical features with inherent missingness. The predictive performance is compared to that of a related deep generative baseline.

Significance

Phenotyping is a persistent problem in biomedical informatics. In most cases, the problem reduces to classification: we are merely interested in identifying which patients are cases for a given clinical feature of interest. This framing suggests the use of supervised methods for solving the phenotyping problem. However, clinical data typically lack the characteristics needed for effective supervised learning. First, these data routinely contain significant missingness. This is a byproduct of how the data are generated; not all clinical variables

are measured for every patient at every encounter. Unfortunately, supervised learning algorithms commonly assume the data are fully observed. More importantly, clinical data suffer from a paucity of high quality, gold standard labels. This is also a byproduct of how the data are generated; they are produced as a means of documentation, not research. The work proposed herein attempts to address these realities by developing phenotyping methods which assume the data, the labels, or both are only partially observed.

Contributions

Aim 1. Implement a method for learning phenotypes from partially observed data. Multi-Channel Mixed Membership Models (MC3M) represent an unsupervised approach to phenotyping, which model different, but equally important sets of observations in their own "data channels". In this aim, we use MC3M to learn phenotypes from survey data designed to assess a wide variety of mutable and immutable factors underlying the health of individuals. Modeling the data associated with each such factor in their own separate channel gives all factors a fair opportunity to influence the structure of learned phenotypes and to be noticeably represented within them. This property is significant since both mutable and immutable factors play important roles in behavior, but only mutable factors offer targets for intereventions. MC3M offers an opportunity to learn phenotypes which take both into account and present them interpretably, on equal footing.

Aim 2. Develop a method for learning phenotypes from partially labeled clinical data. Semi-Supervised Mixed Membership Models (SS3M) introduce a novel mechanism for semi-supervision within a class of models developed for inferring phenotypes from multimodal clinical data. In this way, SS3M permits the researcher to specify which phenotypes they would like the model to learn. Importantly, this added input is minimal: only a subset of cases need to be labeled as positive or negative for the model to learn a disease-specific phenotype. Moreover, SS3M's phenotypes are straightforward to evalute both qualitatively

and quantitatively. Since SS3M phenotypes are easily visualized, a domain expert may visually inspect them to determine how well they represent a targeted disease. Meanwhile, because SS3M performs multi-label classification, the model may be evaluated quantitatively with the usual evaluation metrics for supervised models.

Aim 3. Develop a method for learning phenotypes from partially observed, partially labeled data. Variational Autoencoders (VAEs) are a popular and powerful framework for building deep generative models. However, many VAEs are not well suited to modeling clinical data due to their missingness and label paucity. Our Monte Carlo Expectation-Maximization (MCEM) algorithm addresses these issues by allowing VAEs to train on partially observed data containing a heterogeneous mixture of continuous and discrete values. By addressing both issues simulataneously, our approach facilitates the use of VAEs for clinical classification tasks, in particular supervised disease phenotyping.

Chapter 2

Background and Related Work

2.1 Historical Background

The present work focuses on *clinical phenotyping*. We have yet to succinctly define this term, and, in fact, a consensus definition has eluded the biomedical informatics community (Shivade et al. 2014; Banda et al. 2018). Here we will explore the evolution of topics which undergird clinical phenotyping. Our aim is to provide an understanding of how and where clinical phenotyping fits in this more general context.

As the name implies, clinical phenotyping is a specialization of a larger topic, namely "phenotyping". To better understand clinical phenotyping, it will be helpful to gain a handle on phenotyping. To do that, of course, we may begin by exploring what is meant by the term "phenotype".

Of Phenotypes and Genotypes. Typically, one first encounters the term "phenotype" in the context of genetics where it often appears alongside the term "genotype." In *Molecular Biology of the Gene*, James D. Watson provides concise, familiar definitions for both: "We refer to the appearance (physical structure) of an individual as its phenotype, and to its genetic composition as its genotype." (Watson 1970)

This clean distinction between phenotype and genotype may seem obvious today, but it was not always so. In the late 19^{th} and early 20^{th} centuries, during the early days of genetics, practitioners routinely mixed phenotypic and genotypic concepts (Mayr 1973). Major disagreements among the various schools of thought at the time — Darwinians, Mendelians, and

biometricians most prominently — have been attributed to this confusion. Thus, the failure to separate phenotype from genotype is thought to have delayed by decades the "modern synthesis" which unified the field and formed the basis of modern genetics (Mayr 1973).

In 1909, during the pre-synthesis era, the Danish geneticist Wilhelm Johannsen originally coined the terms "phenotype" and "genotype" (as well as the term "gene") in the first edition of his text, *Elemente* (Churchill 1974). Unfortunately, his attempt to establish these concepts was muddled; we would have to wait until the third edition of Johannsen's text in 1926 to obtain definitions familiar to a reader today:

> **phenotype:** The word phenotype ... can simply be used as a designation of personal charaters of any individual whatever. The phenotype of an individual is thus the embodiment of all of his expressed characters. The single organism, the individual plant, an animal, a man, — "What it is and what it does" — has its phenotype, i.e., it appears as a sum of traits which are determined by the interplay between "inherited [genes]" and elements of the environment. (Churchill 1974)

> **genotype:** The basis for the entire development of an individual is ... given by the constitution of the two gametes, by the union of which the organism arises. This constitution we thuse designate with the word genotype. (Churchill 1974)

The Extended Phenotype. In the years and decades following Johannsen's *Elemente* the genotype-phenotype division was firmly integrated into genetic theory. Though the defintions continued to evolve, the core of the original phenotype concept remained intact: an organism's observable traits correspond to its phenotype. In 1982, Richard Dawkins introduced a novel perspective which expanded the domain of the phenotype beyond the physical limits of the organism. His "extended phenotype" broadened phenotypes to include the effects organisms imposed on their environments (Dawkins 1982). That is, the pond a

beaver creates by damming a river is a much a part of its phenotype as the color of its fur or the width of its tail.

Phenotypes in Phenomes. In the last several decades various fields of study have emerged sharing the suffix "-omics": genomics, transcriptomics, proteomics, metabolomics, etc., each focused on the collection and study of large quantities of phenotypic data. From this literature, the idea of the phenome and its study — phenomics — has emerged (Freimer and Sabatti 2003). The phenome is meant to connote the structured totality of an organism's phenotypic content. As such, the term is somewhat redundant to "phenotype." However, it does add a sense of organization. C. R. Scriver nicely illustrates the concept:

> The phenome comprises layer upon layer of phenoytpes; it exists as compartments within compartments in cellular and organismal anatomy. The ... organism comprises molecules inside organelles inside cells inside organs, all enclosed with an integument, all connected by plumbing, wiring and telecommunications, the parts sending and receiving momnet-to-moment information; all resident in the fluctuating traffice of temporal experience. (Scriver 2004)

The Digital Phenotype. The phenome communicates the idea that phenotypes can be organized into intercommunicating phenotypic layers. The extended phenotype, adds to these layers some that lay outside the physical limits of the organism. The *digital phenotype* (Jain et al. 2015; Loi 2019; Coghlan and D'Alfonso 2021) simply adds another extra-organismal phenotypic layer to the phenome, a layer accessed via interactions with technology (Coghlan and D'Alfonso 2021). Michele Loi defines the term as inherently human-centric:

> Hence, I propose to characterize the human digital phenotype as *an assemblage of information in digital form, that humans produce intentionally or as a by-product of other activities, and which affects human behavior.* More succinctly (but less

> precisely), the human digital phenotype consists of digital information *produced by humans* and *affecting humans*. (Loi 2019)

With the introduction of digital phenotypes, the phenotype — or equivalently, the phenome (Mahner and Kary 1997) — concept has been updated to accommodate the full spectrum of (human) organism traits in modern times.

From Phenotype to Phenotyping. In general, "phenotyping" refers to task of identifying individual organisms within a species which share a common phenotype (Shivade et al. 2014; Banda et al. 2018; Li, Zhang, and Huang 2014; Furbank and Tester 2011). In most cases, the targeted "phenotype" is charecterized not by all observable traits, but rather by a relatively, small manageable subset (Mahner and Kary 1997). For example, commercially grown soy plants may be phenotyped according to leaf or root morphology or their total yield (Li, Zhang, and Huang 2014; Furbank and Tester 2011). More relevant for our purposes and as we discuss next, people may be phenotyped based on the digital artifacts they create while interacting with technology.

Digital Phenotyping. People interface with technology in myriad ways on a daily basis. Some of these technologies (e.g. personal computers, mobile phones, wearables), are constantly collecting data that the user generates both actively and passively. "Digital phenotyping" refers to the identification of individuals expressing common digital phenotypes within these data streams (Jain et al. 2015). Noteably, much of the literature on digital phenotyping has focused on health, in particular mental health. Nevertheless, reliance on digital technologies is ubiquitous, and so the term may be applicable more broadly.

(Digital) Clinical Phenotyping. We now have the context and concepts to define the term "clinical phenotyping." Clinical phenotyping is a type of digital phenotyping; it is used to identify patients who express a common digital phenotype in digital health data such as

electronic health records (EHRs) (Shivade et al. 2014; Banda et al. 2017). Most often, a clinical phenotype is meant to herald the presence, within a patient, of a specific disease of interest or an exposure to a specified clinical intervention (Banda et al. 2017). Novel phenotypes may also be uncovered which may suggest disease subtypes.

2.2 Literature Review

Various methods have been used for clinical phenotyping. We review these here.

Rule-Based Methods. The use of rules to filter a patient population constitutes the most traditional and most common approach to clinical phenotyping. The process usually begins with a target (e.g. a disease or clinical intervention) and, in collaboration with clinical experts, then proceeds to iteratively refine a set of inclusion and exclusion criteria for constructing a cohort of cases. Commonly, these criteria are designed to operate on structured clincal data types such as lab values, medications, procedures, diagnosis codes, and other metadata (Banda et al. 2018). For example, using note metadata, Essay, Mosier, and Subbian 2020 constructed a phenotyping decision tree for identifying cases for any one of seven distinct interventions applied in the context of respiratory failure.

Peformant rule-based phenotypes using only diagnosis codes have been developed for many conditions including autoimmune disorders (Nicholson et al. 2013), pediatric metabolic disorders (Lingren et al. 2016), and cardiovascular diseases (Fan et al. 2013; Morley et al. 2014; Esteban et al. 2017a). Similarly, performant phenotypes for cardiac interventions have been built using just procedure codes (Petersen et al. 1999). However, phenotypes may often benefit from inclusion of multiple data types. For example, Schmiedeskamp et al. 2009 found that phenotypes for *Clostridium difficile* infections performed best when containing both diagnosis codes and medication data.

The rule-based approach to phenotyping is so commmon, that tools have emerged for facilitating their development and validation (Xu et al. 2015). For example, the Observa-

tional Health Data Sciences and Informations (OHDSI) collaborative have made their cohort construction tool, ATLAS, publically available (Hripcsak et al. 2019). When paired with clinical data formatted according to the Observational Medical Outcomes Partnership (OMOP) common data model (Overhage et al. 2012; Hripcsak et al. 2015), ATLAS may be used to implement rule-based queries and visualize cohort statistics in near real time.

Creating and validating rule-based phenotypes can be a difficult enterprise (Kirby et al. 2016). Thankfully, the biomedical informatics community has worked to make validated phentoypes publically available. The Electronic Medical Records and Genomics (eMERGE) network maintains a catalog of many phenotypes covering various conditions, drug responses, and other clinical concepts (Kho et al. 2012; Denny et al. 2011; Ritchie et al. 2010). These are made available through the Phenotype Knowledge Base (PheKB), a repository which collects phenotyping algorithms from many sources (Kirby et al. 2016).

Rule-based phenotyping methods are straightforward conceptually, and can be relatively simple to implement with modern tools; However, they have several limitations. First, it can be difficult and time-consuming for experts to reach consensus on phenotyping criteria (Newton et al. 2013). This bottleneck is exacerbated if several rounds of criteria refining are necessary. Exporting rule-based algorithms to other datasets may also be a challenge (Bayley et al. 2013). Rules are usually encoded as database queries specialized to run on data formatted according to specific data model, and thus cannot be easily applied to data formatted differently. For this reason, rule-based phenotypes are often shared in descriptive formats including text, figures, and pseudocode. The lack of phenotype interoperability across data models is a major driver behind the development and adoption of common data models like those maintained by OHDSI (Hripcsak et al. 2015), Patient-Centered Clinical Research Network (PCORnet) (Califf 2014), Informatics for Integrating Biology and the Bedside (i2b2) (Murphy et al. 2010), and Mini-Sentinel (McGraw, Rosati, and Evans 2012). Finally, rule-based phenotypes are informed but also constrained by expert clinical knowledge. Since they encode clinical guidelines and clinical experience, they will miss potentially

helpful correlations that may be detected and exploited by more data-driven methods.

Machine Learning. Instead of relying on experts to distill their knowledge into phenotyping rules, many authors use machine learning methods to learn phenotyping algorithms directly from data. These methods may be grouped according to their reliance on labels; they supervised, unsupervised, semi-supervised, inaccurately supervised methods; we explore each group below.

Supervised Methods. When the data contain a full complement of labels encoding the presence or absence of the target of interest, supervised methods are a natural choice for phenotyping. In this setting phenotyping is reduced to classification: the objective is to obtain a discriminative model which approximate $P(Y|X)$ — the probability of the target labels, Y, conditional on the features, X.

Many supervised models have been utilized for phenotype learning including logistic regression (Liao et al. 2010; Heintzelman N.H. et al. 2013; Kumar V. et al. 2014; Kamkar et al. 2015; Bhattacharya M. et al. 2017; Koola J.D. et al. 2017; Kummer B.R. et al. 2017; Zheng et al. 2017; Rotmensch et al. 2017; Gustafson et al. 2017; Geva et al. 2017; Blecker et al. 2017; Koola et al. 2018), random forest (Kamkar et al. 2015; Zhou et al. 2016; Bhattacharya M. et al. 2017; Teixeira P.L. et al. 2017; Chaganti et al. 2017; Turner et al. 2017; Koola et al. 2018), support vector machines (Wei et al. 2010; Carroll, Eyler, and Denny 2011; Kotfila and Uzuner 2015; Zheng et al. 2017; Kagawa et al. 2017; Turner et al. 2017; Koola et al. 2018), naive bayes (Zheng et al. 2017; Rotmensch et al. 2017; Turner et al. 2017; Koola et al. 2018; Orphanou et al. 2018), and neural networks (Che et al. 2015a; Che et al. 2015b; Lipton, Kale, and Wetzel 2015; Geraci et al. 2017; Gehrmann et al. 2018) (see below). These models are commonly trained using data extracted from structured EHR data fields. To boost performance researchers may add features extracted from unstructured notes using natural language processing (NLP) engines (Liao K.P. et al. 2015; Lin et al. 2015) such as the Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) (Savova et al.

2010) or MetaMap (Aronson 2001; Aronson and Lang 2010).

Supervised phenotyping methods may also be developed to accomodate temporal information. Lin et al. 2015 extract temporal features from longitudinal clinical time-series using a custom NLP module. These features are then combined with structured and other NLP-derived features and aggregated across the time-dimension to yield vectors consumed by a penalized logistic regression model used to predict methotrexate-induced liver injury. Lipton, Kale, and Wetzel 2016 use Recurrent Neural Networks (RNNs) to model patient trajectories. The RNN hidden states are then used for multilabel prediction of 128 condition labels derived from diagnosis codes. Similarly, Harutyunyan et al. 2019 use RNNs for phenotyping 25 acute care conditions. Bejan et al. 2013 use features extracted from clinical reports to phenotype pneumonia *throughout* patients' inpatient trajectories. Note, this was possible due their use of expert annotators who reviewed each report in each patient time-series for presence of pneumonia.

Supervised machine learning algorithms share a common, major limitation: both the training and validation data must be fully labeled. As such, most of the studies cited above required extensive chart review by multiple clinical experts. Otherwise, researchers may choose to model the raw diagnosis codes (Lipton, Kale, and Wetzel 2016; Miotto et al. 2016a) which may not accurately reflect patient state (Hogan and Wagner 1997) or use labels generated using validated phentoyping algorithms as ground truth (Ding et al. 2019). The need for expert review is a severe bottleneck that restricts the scalability of supervised phenotyping in most circumstances. Later, we will review methods that attempt to alleviate this constraint.

Unsupervised Methods. In the absence of reliable labels, unsupervised methods may be used to identify similar subsets of patients. Patients in each subset may be understood to be express a common phenotype.

Tensor factorization and mixed membership models are used to learn the structure of

latent phenotypes from the data (Ho, Ghosh, and Sun 2014a; Ho et al. 2014a; Wang et al. 2015a; Hui et al. 2015; Perros et al. 2015; Perros et al. 2017; Henderson et al. 2017; Kim et al. 2017; Ruffini, Gavalda, and Limon 2017). In addition, these models learn low-dimensional patient representations which encode how well each latent phenotype "explains" each patient's observations. Hierarchical clustering (Tamang and Parsons 2011; Marlin et al. 2012; King E. et al. 2014; Doshi-Velez, Ge, and Kohane 2014; Chubachi et al. 2016; Levoska M. et al. 2017) and k-means/medioids (Hong S.B.N. et al. 2015; Vazquez Guillamet et al. 2016; Cerna A.E.U. et al. 2017) use distance metrics to calculate patient similarity and group similar patients together. Probabilistic graphical models (Marlin et al. 2012; Tran et al. 2015; Schulam, Wigley, and Saria 2015; Russo C. et al. 2017; Mayhew et al. 2018) utilize latent variables to cluster similar patients, while autoencoders (Lasko, Denny, and Levy 2013; Suresh, Szolovits, and Ghassemi 2017) use hidden layers in an analagous way. Word embedding algorithms learn are used to learn low-dimensional vector representations of clinical concepts which can be aggregated to generate embedded representations of patients; proximity of patients to disease concepts can be used to phenotype (Gligorijevic, Stojanovic, and Obradovic 2016; Glicksberg et al. 2018).

Several unsupervised phenotyping methods explicitly handle time. Tensor factorization methods that model a temporal axis have been used to inform the structure of latent phenotypes (Zhou et al. 2014; Perros et al. 2017; Zhao et al. 2019). Custom probabilistic graphical models have been developed to learn phenotypes from physiologic time-series (Saria, Koller, and Penn 2010; Saria, Duchi, and Koller 2011; Schulam, Wigley, and Saria 2015). RNNs which model time-series naturally, have been employed to phenotype disease subtypes (Che et al. 2017; Xu et al. 2020).

Relative to supervised phenotyping methods, unsupervised methods avoid the costs of expert annotation; however, evaluation is often more complex. This is because unsupervised methods oftern function by identifying clusters of similar patients, but these clusters may or may not correspond to recognizable phenotypes.

Semi-Supervised Methods. In semi-supervised learning, the data are assumed to be only paritally labeled. That is, only a subset of the training instances are paired with labels. As such, semi-supervised learning occupies a space between supervised and unsupervised learning (Zhou 2018). When using these methods, the goal is to leverage information about the unlabeled data distribution, $P(X)$, to obtain a better estimate of $P(Y|X)$ (Chapelle, Schölkopf, and Zien 2006).

In the context of phenotyping, semi-supervised methods can limit the expense of expert annotation, while still retaining some of the major benefits of supervised phenotyping: labels specifying the target to phenotyped and ease of evaluation. As such, semi-supervised phenotyping has been proposed by some authors though not as many as have explored supervised and unsupervised phenotyping. Garla, Taylor, and Brandt 2013 employed Laplacian SVMs to detect malignant liver lesions in imaging reports. Roqueiro et al. 2015 use co-training to distinguish between patients suffering from either of two classes of migraine. Dligach, Miller, and Savova 2015 use expectation-maximixation to learn phenotypes for chronic conditions including ulcerative colitis, crohn's disease, multiple sclerosis, and type II diabetes. Henderson et al. 2018 leverage semi-supervised tensor factorization to learn phenotypes for type II diabetes and resistant hypertension. Finally, Zhang et al. 2019 introduce PheCAP, a semi-supervised pipeline for learning phenotypes which relies on unsupervised feature learning and modeling a small set of expert-generated, gold standard labels using supervised classifiers.

Inaccurately Supervised Methods. Inaccurate supervision is a type of supervised learning in which the labels are assumed to contain errors (Zhou 2018). In the context of phenotyping, this type of learning relies upon imperfect, noisy labels which can be easily and cheaply extracted in large quantities from structured data fields (e.g. diagnosis codes) or text using NLP pipelines.

In a series of studies, Halpern *et al.* introduced the idea of anchors — variables that herald a true state Y and which are conditionally independent of all other variables given Y — and

used them as predictive targets to learn phenotypes using supervised classifiers (Halpern et al. 2014; Halpern, Horng, and Sontag 2016; Halpern et al. 2016). In a similar fashion, (Agarwal et al. 2016; Yu et al. 2017; Wagholikar et al. 2020) employed easily extracted silver-standard labels to train supervised phenotyping algorithms. The popularity of this approach even led to an automated phenotyping software package designed for the OMOP common data model (Banda et al. 2017). Subsequent studies incorporated more elaborate feature engineering to construct better, more interpretable classifiers trained with silver standards (Yu et al. 2018; Ahuja et al. 2020). Silver-standards have also been incorporated into a temporal phenotyping algorithm for detecting onset of chronic and acute conditions in clinical time-series (Ferté et al. 2021).

2.3 Review of Literature Gaps

In reviewing the literature, there emerged several desiderata for phenotyping algorithms:

(i) **Target specification.** Phenotyping algorithms are most commonly developed for identifying patients who are likely cases of known clinical targets. As such, incorporation of a target specification mechanism (e.g. supervision) is often needed to construct a useful phenotyping algorithm.

(ii) **Minimizing chart review.** Gold-standard labels are often needed to evaluate phenotyping algorithms; they also may be needed for training. Generating such labels requires expert chart review which is difficult, time-intensive, and limits scalability. As such, phenotyping algorithms should minimize reliance on gold-standard labels as much as possible.

(iii) **Robustness to missingness.** Clinical data commonly suffer from missingness in features and in labels. This missingness reflects the nature of clinical data: not all variables are measured at all times for all patients (feature missingness), and the data

are meant for documentation, not research (label missingness). Phenotyping algorithms should recognize this reality and take feature and label missingness into account.

(iv) **Interpretability.** In the clinical domain, model interpretability is often desired: models should explain themselves if they're going to influence decision making. Similarly, interpretable phenotyping algorithms are often preferred over uninterpretable ones.

Many existent phenotyping methods address (i) and (ii); however, there is a lack of methods that handle (iii) and (iv). Aim 1 of this dissertation addresses (ii-iv) with an unsupervised, interpretable probabilitic graphical model which is robust to missingness. Aim 2 addresses (i-iv) incorporating semi-supervision into the model utilized in Aim 1. Finally, Aim 3 addresses (iii) with an algorithm for training deep generative models on partially observed data, and could be extended to cover (i) and (ii) by extending the algorithm to related semi-supervised models.

Chapter 3

Aim 1. Implement a method for learning phenotypes from partially observed data.

When subjected to an intervention, different individuals will often experience different responses. Such heterogeneity in treatment effect has been observed for various behavioral interventions targeting a wide range of health outcomes (Burgermaster et al. 2017; Koch, Contento, and Gray 2019; Rothman and Sheeran 2020; Bryan, Tipton, and Ds 2021). Much of this variability may be attributed to differences in individuals' underlying psychological and social context. Thus, methods for identifying and characterizing these differences could facilitate a transition from merely describing treatment effect heterogeneity to optimizing treatment effect by tailoring interventions to better suit an individual's context.

Phenotyping methods generally aim to identify subgroups of individuals who share observable similarities due, presumably, to a set of common, underlying factors (Dugger, Platt, and Db 2018). Such methods are familiar tools in genetics and medicine where they have contributed to the development of precision health — a field which aims to tailor medical interventions to the specific biological and medical context of individual patients (Collins and Varmus 2015). In the present work, we aim to extend the use of phenotyping methods to identify subgroups with similar psychological, social, and environmental characteristics, that is subgroups characterized by distinct "psychosocial-behavioral phenotypes". We do this in an effort to prospectively identify the potential behavioral, psychological, social, and environmental factors which may be responsible for variability in response to a potential health behavior change intervention.

To this end we aim to construct psychosocial-behavioral phenotypes which 1) comprise psychosocial determinants of health (e.g., emotions, attitudes, mental health, self-esteem, social support integration, environmental stress, discrimination, among others (Matthews, Gallo, and Se 2010)); 2) represent distinct, conceptually meaningful subgroups; 3) highlight mutable factors that can be intervened upon (i.e. mediators) as well as immutable contextual factors (i.e. moderators); and 4) are related to health outcome of interest (Hickey, Bakken, and Byrne 2019; Kim et al. 2021). We take special care to note that while there does exist prior work in this space (Fuentes, Brondeel, and Franco 2019; Boutelle et al. 2014; Bouhlal et al. 2017; Burgermaster et al. 2018), none so far have modeled contextual factors alongside potential intervention targets.

Machine learning methods developed for clinical phenotyping (Pivovarov et al. 2015a) may be applicable to this problem. Specifically, Mixed Membership Models, a model family which includes Latent Dirichlet Allocation (LDA), a popular method from the topic modeling literature (Blei, Ng, and Jordan 2003), could be used to identify psychosocial-behavioral phenotypes. Here we apply Multi-Channel Mixed Membership Models (MC3M), a relative of LDA which permits modeling of multiple data types in separate "channels." In clinical phenotyping, MC3M's channels have previously been used to accommodate clinical note text, laboratory tests, medications, and diagnosis codes (Pivovarov et al. 2015a; Lu, Wei, and Hsiao 2016). In our work, MC3M's channels are allocated to survey responses relating to distinct psychosocial-behavioral constructs associated with health behaviors. By modeling multiple constructs together, we aim to generate phenotypes comprising behaviors and psychosocial determinants of health that include both mediators and moderators. Modeling these constructs jointly within a single model has the capacity to identify both potential intervention targets and important contextual factors within previously unknown population subgroups.

In this work, we apply MC3M to survey data gathered to investigate how psychosocial determinants of health are associated with health status within an inner-city community

burdened with health disparities including high levels of overweight and obesity. Next, using predictive modeling, we identify the subset of MC3M phenotypes which carry a significant association to elevated weight status (BMI $\geq 25\text{kg/m}^2$) — an important risk factor for chronic disease driven by complex interactions among psychological, behavioral, social and environmental determinants (Swinburn, Sacks, and Kevin D. Hall 2011; Calugi and Dalle Grave 2020). Our goal here is to demonstrate how MC3M can be used to proactively identify contextual differences within a target population which could be subsequently incorporated within a behavioral intervention design. We aimed first to identify unique subgroups of participants with similar psychosocial characteristics that highlight potential intervention targets and important contextual conditions (i.e., psychosocial-behavioral phenotypes). Then we aimed to assess the potential utility of these phenotypes for intervention design by determining if 1) phenotypes were differently distributed among people in our dataset and 2) only some phenotypes were related to a relevant health outcome, weight status, in our dataset.

3.1 Aim 1A. *Utilize the Multi-Channel Mixed Membership Model (MC3M) to learn psychosocial-behavioral phenotypes from partially observed survey data*

Background

Treatment effect heterogeneity is routinely observed in response to interventions meant to improve health outcomes. This heterogenity is likely contributed to by differences in psychological and social context among treated individuals. Thus, identifying and characterizing these differences via psychosocial-behavioral phenotyping could improve intervention design by taking contextual information into account. Here we utilize Multi-Channel Mixed Membership Models (MC3M) to learn psycho-social phenotypes from survey data. First, we describe the survey response dataset we use throughout the present aim. Then we detail the structure of MC3M's probabilistic graphical model as well as an inference algorithm which

allows the model to learn phenotypes from partially observed data. Finally we train MC3M on tokenized survey response data and inspect the extent to which MC3M's phenotypes identify distinct subgroups within the survey sample population.

Research Question

Can MC3M identify subgroups of individuals that exhibit similar psycho-social traits?

Methods

Data. The Washington Heights Informatics Infrastructure for Comparative Effectiveness Research (WICER) project was created to improve health equity within immigrant communitys in Manhattan's Washington Heights and Inwood neighborhoods (Lee et al. 2015; Lor et al. 2019; Masterson Creber et al. 2017; Sepulveda-Pacsi and Bakken 2017; Yoon, Suero-Tejeda, and Bakken 2015). WICER participants were recruited from community households and ambulatory care clinics affiliated with NewYork Presbyterian Hospital – Columbia University Medical Center between 2010 to 2013 (Lee et al. 2015; Lor et al. 2019). Data were collected for 5,883 adult participants who were 18 years of age or older (Masterson Creber et al. 2017).

This work, uses the WICER Community Health Surveys, which contains each participant's responses to questions about health behaviors, healthcare interactions, health, mental health, physical activity and diet, sleep, stress, social network, attitudes, neighborhood, and health literacy. Detailed information about the questionnaire was previously published (Lee et al. 2015). We include surveys that assessed psychosocial and behavioral constructs potentially relevant to health behavior intervention design. We grouped questions according to the construct they were developed to target (e.g., diet, neighborhood food environment, stress). Table 3.1 summarizes the constructs targeted by surveys included in WICER.

WICER also includes height (meters [m]) and weight (kilograms [kg]) measurements. From these, we calculated each participant's body mass index (BMI). A BMI $\geq$ 25kg/m2 is

a positive indicator for elevated weight status (i.e., overweight or obesity). We encoded each participant's weight status as a binary variable reflecting BMI $<$ 25kg/m2 or $\geq$ 25kg/m2.

MC3M is a model for tokenized data. To apply MC3M, we translated the WICER survey data into survey response tokens which are unique combinations of question-answer pairs (e.g., `Q:FruitVegetablesAvailableInNeighborhood_A:Somewhat`) or validated composite scale scores (e.g. `perceivedStressScore_moderate`). These tokens were then manually translated to be more easily interpreted (e.g. `perceivedStressScore_moderate` became `some stress`). Preprocessed survey data and weight status indicators were aggregated to form our full preprocessed datasets which we then split into training (80%), validation (10%), and test (10%) partitions.

Model. Let D, C, and P be the total number of persons, constructs, and phenotypes respectively. Let W represent our complete dataset of discrete observations. We may subdivide W into observations associated with each person: $W = \{W_1, \ldots, W_D\}$. Furthermore, each person's observations, W_d, may be subdivided by construct: $W_d = \{W_{d1}, \ldots, W_{dC}\}$. Finally, each W_{dc} may be further decomposed into a set of individual observations: $W_{dc} = \{w_{dc1}, \ldots, w_{dc(N_{dc})}\}$, where N_{dc} is the total number of observations of construct c associated with person d.

MC3M assumes the existence of a set of phenotypes, $\Phi = \{\phi_1, ..., \phi_P\}$, each of which has a component dedicated to each construct: $\phi_p = \{\phi_{p1}, ..., \phi_{pC}\}$. Phenotypes are modeled as a set of C independent discrete probability vectors. Each of these vectors is a discrete probability distribution over the corresponding construct's survey response tokens. These phenotype-token distributions are sampled from a set of Dirichlet distributions with fixed parameters, $\beta = \{\beta_1, ..., \beta_C\}$.

$$P(\Phi; \beta) = \prod_{p=1}^{P} \prod_{c=1}^{C} \mathrm{Dir}(\phi_{pc}; \beta_c) \tag{3.1}$$

MC3M also assumes a set of person-specific distributions over phenotypes or person-

phenotype distributions: $\Theta = \{\theta_1, ..., \theta_D\}$. These distributions are modeled as samples from a single Dirichlet distribution with fixed parameter α.

$$P(\Theta; \alpha) = \prod_{d=1}^{D} \mathrm{Dir}(\theta_d; \alpha) \tag{3.2}$$

In MC3M, the observations in W are assumed to be generated by interactions among the probability vectors in Θ and Φ. To illustrate, consider the observations for a single person, d' and construct c', $W_{d'c'}$. Each individual observation $w_{d'c'n}$ in $W_{d'c'} = \{w_{d'c'1}, ..., w_{d'c'(N_{d'c'})}\}$ is modeled as a sample from the $c'^{\text{ th}}$ component of one of the phenotypes in Φ. The identity of this phenotype is obtained by sampling a phenotype assignment, $z_{d'c'n}$ from $\theta_{d'}$, where $z_{d'c'n} \in \{1, ..., P\}$. Thus, MC3M assumes that each observation in W is obtained by first sampling an assignment from a person-phenotype distribution in Θ, picking out the assigned phenotype from Φ, and finally sampling the observation from one of the phenotype-token distributions. This allows us to write out the conditional probabilities of all the assignments, Z, and observations, W, as follows.

$$P(Z|\Theta) = \prod_{c=1}^{C} \prod_{d=1}^{D} \prod_{n=1}^{N_{dc}} \theta_{d(z_{cdn})} \qquad P(W|\Phi) = \prod_{c=1}^{C} \prod_{d=1}^{D} \prod_{n=1}^{N_{dc}} \phi_{(z_{cdn})c(w_{cdn})} \tag{3.3}$$

Taken together, the probability distributions in equations 3.1, 3.2, and 3.3 fully specify the generative model for MC3M.

Inference. Here we describe the inference algorithm we implement to obtain estimates of the variables Θ and Φ given our training data, W.

We adopt a Bayesian inference approach, and seek to obtain posterior estimates of Θ and Φ. Unfortunately, these posteriors require we estimate the marginal likelihood of our data, which is intractable. Therefore, we make use of Markov chain Monte Carlo approximate inference methods, which do not necessitate estimation of this marginal likelihood. Specifically, we implement a collapsed Gibbs sampling algorithm for our model as decribed by Pivovarov et al. 2015a.

Briefly, we integrate out of the model's joint distribution the latent variables Θ and Φ. We then iteratively sample each of the the assignment variables in Z from its complete conditional distribution which is the distribution of the assignment variable conditioned on the values of all other remaining variables in the model. We repeat this sampling procedure until observing convergence in the model's collapsed likelihood. At this point, we may recover the values of Θ and Φ as Monte Carlo estimates of their expectations with respect to the collapsed likelihood.

Model Selection. To optimize the total number of phenotypes, P, we carry out three independent Gibbs sampling runs for a range of values: 10, 15, 20, 25, 30, 35, and 50. We utilize a Chib-style estimator (Wallach et al. 2009; Murray and Rr 2008) to estimate the held-out posterior likelihood of our validation set under each model. We then select the value of P that yields the maximum average held-out likelihood as the number of phenotypes for this study.

Identifying informative constructs. To be useful, MC3M should learn phenotypes which differ in how they distribute probability mass over survey response tokens. Moreover, different phenotypes should "emphasize" distinct constructs, because it is probable that only a subset of constructs will be relevant to each phenotype. To identify such constructs, we calculated the entropy of their corresponding phenotype-token distributions. Low entropy constructs have distributions which peak over a relatively small number of tokens. Conversely, high entropy constructs have nearly uniform distributions which indicate the construct as a whole is irrelevant to the phenotype. We normalized each construct's entropy to compare entropies across distributions with varying numbers of tokens. Normalized entropy was calculated by dividing the entropy by the maximum possible entropy. For a construct containing V survey response tokens, the maximum entropy is equal to $\log V$.

Active and inactive phenotypes. In MC3M, each person in the dataset is described to some extent by every phenotype. Ideally, an individual's person-phenotype distribution should place high probability upon only a small number of phenotypes that explain the large majority of the individual's observations. To identify these active phenotypes, we set a threshold of one standard deviation above the mean of the person-phenotype prior distribution. Phenotypes with probability above this threshold were considered active; otherwise, they were inactive.

Visualizing Phenotypes. We developed a text-based visualization to interpret the information captured by a phenotype. For a given phenotype, we first identified the 20 constructs whose phenotype-token distributions had the lowest normalized entropy. For each such construct, we then selected all survey response tokens with probability at least equal to the phenotype-token distribution's prior mean probability plus 50% of the prior standard deviation. For each such token, we calculated a relative probability equal to the token probability divided by the maximum token probability under the present phenotype and construct's phenotype-token distribution. As in previous related work (Pivovarov et al. 2015a; Rodriguez and Perotte 2019), we visualized these tokens with word clouds in which (1) token font sizes were proportional to their relative probabilities and (2) font colors were unique to each construct. This visualization is useful for interpreting phenotypes because it illustrates the variation in relative probabilities within each phenotype-token distribution. Since even the largest token probabilities within a phenotype-token distribution can be quite small when the total number of tokens is large, relative probabilities communicate the most interpretable information captured by each phenotype.

Results

Sample Population The mean age of participants was 50 years (SD=17) and nearly three-quarters of participants were female. Nearly all participants (96%) reported Hispanic

ethnicity and two-thirds completed the surveys in Spanish. Most participants were insured by Medicaid and had an elevated weight status (BMI $\geq$ 25 kg/m2).

Phenotype Learning Models trained with $P = 20$ obtained the maximum held-out log-likelihood averaged over all runs (Figure 3.1). Therefore, we used MC3M person-phenotype and phenotype-token distributions learned, setting P=20 in all subsequent analyses.

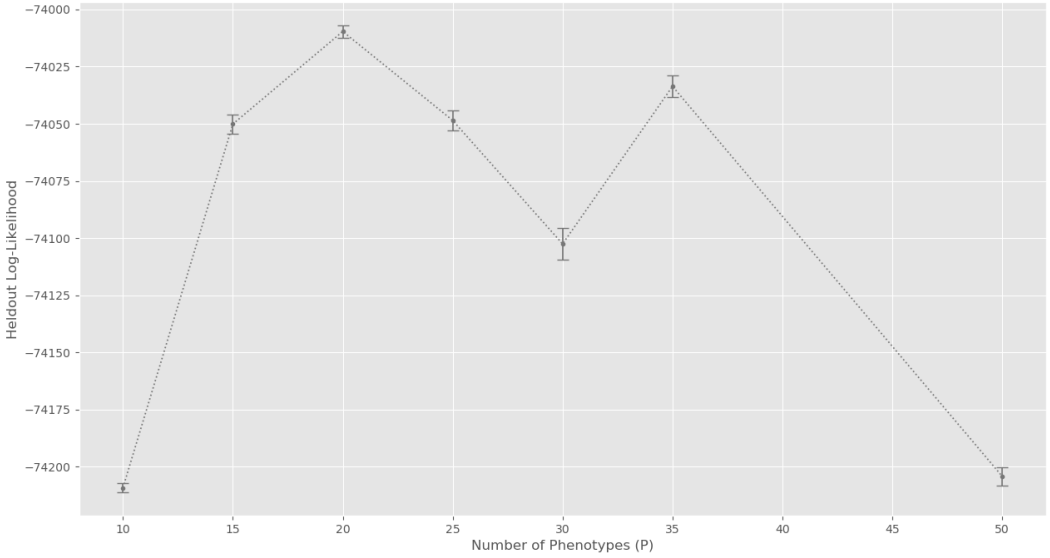

Figure 3.1: Model selection. Shown are held-out posterior likelihood estimates of the validation set data under for $P \in \{10, 15, 20, 25, 30, 35, 50\}$. Error bars correspond to standard error over mutliple runs of the Chib-style estimator. The maximum value was observed at $P = 20$.

Psychosocial-Behavioral Phenotypes A subset of the 20 psychosocial-behavioral phenotypes is presented in Figures 3.2 and 3.3. The word clouds are presented so that each color represents a different construct (See Table 3.1 for details on each construct). In each word cloud, constructs are ordered by increasing entropy so that the most discriminative constructs are presented first. Within each construct, tokens that contribute the most probability mass to the phenotype are presented with each token's font size corresponding to its

relative probability. Although some phenotypes include tokens in common, the combination of tokens is unique across phenotypes.

Constructs and Measures Included in Psychosocial-Behavioral Phenotypes

Construct	Scale(s)	Tokens	Example Concepts
Diet	NYC HANES[a]	63	Fruit/vegetable intake, beverage intake, restaurant meal frequency
Alcohol Consumption	NHANES[b]	2	Frequency and quantity of alcohol consumption
Smoking	NHANES[b]	2	Cigarette smoking
Physical Activity	NYC HANES[a]	15	Weekly minutes of walking, moderate and vigorous physical activity
Sedentary Behavior	NYC HANES[a]	12	Computer, TV screentime
Sleep	MOS[c], PROMIS[d]	12	Quantity of sleep, napping
Attitudes	Boden[e], CPS[f], MEPS[g], WICER CHS[h]	1772	Health worries, medical skepticism, medical decision making
Outcome Expectations	MEPS[g]	43	Future discounting, life/health expectancy
Health Literacy	NVS[i], Chew[j]	10	Newest Vital Sign score, medical literacy
Health Locus of Control	MHLC[k]	12	Internal/external/powerful others locus of control
Health Information Seeking	NHCS-HIT[l]	10	Use of internet for health information
Self-perception	Donahue[m]	5	Perception of activity/weight compared to peers
Physical Activity Barriers	Donahue[m]	25	Physical activity environment
Social Support	NTC[n], PROMIS[o], Duke SSI[p]	113	Neighborhood perceptions, social relations, social network, social support
Food Environment	Moore[q]	15	Access to vegetables, healthy food options
Chronic stress	chronic burden of stress[r]	9	Financial stress, job stress, health stress, relational stress
Demographics	WICER[h], Marin[s], MacArthur[t], Duke SSI[p]	257	Gender, household, nativity, race, ethnicity, education, language, age
Health Status	HRQOL SF-8[u], CDC HRQOL14[v], BRFSS[w]	91	Self-reported cancer, diabetes, stroke, mental health diagnosis
John Henryism	John Henryism[x]	3	Level of high-effort coping
Mental Health	CES-D[y], PHQ-9[z], PROMIS[aa]	75	Anxiety, depression, sleep disturbance, quality of life, pain
Perceived Stress	PSS[bb]	3	High/mid/low composite perceived stress score
Sleep Quality	HRQOL SF-8[u]	15	Sleep difficulty, satisfaction
Socioeconomic Status	CCHS[cc], NHANES[s], MacArthur[t], Duke SSI[p]	1378	Insurance, occupation, food insecurity, social position, income source
Self-Advocacy	MEPS[g]	38	Willingness to communicate assertively with healthcare system

Table 3.1: Notes: a) New York City Health and Nutrition Examination Survey (Thorpe et al. 2006) b)National Health and Nutrition Examination Survey (Health Statistics et al. 2014) c) Medical Outcome Study Sleep Scale (Ware and Sherbourne 1992) d) Patient Reported Outcomes Measurement Information System (PROMIS-sleep disturbance) (Buysse, Yu, and Moul 2010) e) Boden-Albala et al. 2011 f) Control Preferences Scale (Degner, Sloan, and Venkatesh 1997) g) Medical Expenditure Panel Survey, Household Component (J. 1997) h) WICER Community Health Survey (Yoon, Wilcox, and Bakken 2013) i) Newest Vital Sign (Weiss, Mays, and Martz 2005) j) Chew, Griffin, and Partin 2008 k) Multidimensional Health Locus of Control (Ka 2005) l) Health Information National Trends Survey (Nelson, Kreps, and Hesse 2004; Cantor et al. 2009)[REF 67,68] m) Donahue et al. 2004 n) Neighborhood Trust and Cohesion (Garcia, Taylor, and Ba 2007) o) Patient-Reported Outcomes Measurement Information System – Social Role (Hahn, DeVellis, and Bode 2010) p) Duke Social Support Index (Landerman et al. 1989) q) Moore et al. 2008 r) Chronic burden of stress sum score (Yj 2013) s) Short Acculturation Scale for Hispanics (Marin et al. 1987) t) MacArthur Sociodemographic Questionnaire (Lachman and Sl 1998) u) Heath-related quality of life Short Form 8 (Ware et al. 2001) v) Centers for Disease Control and Prevention Health-related quality of life Healthy Days Measures, core and symptom module (*Centers for Disease Control and Prevention Behavioral Risk Factor Surveillance System* 2010) w) Centers for Disease Control and Prevention Behavioral Risk Factor Surveillance System (*Centers for Disease Control and Prevention Behavioral Risk Factor Surveillance System* 2010) x) John Henryism Scale (James et al. 1987) y) Center for Epidemiologic Studies Depression Scale (Eaton et al. 2004) z) Modified Patient Heath Questionnaire-9 (Kroenke et al. 2010) aa) Patient Reported Outcomes Measurement Information System – Emotional Distress (Pilkonis et al. 2011) bb) Perceived Stress Scale (Cohen, Kamarck, and Mermelstein 1983) cc) Canadian Community Health Survey Food Insecurity Measure (Kirkpatrick and Tarasuk 2008)

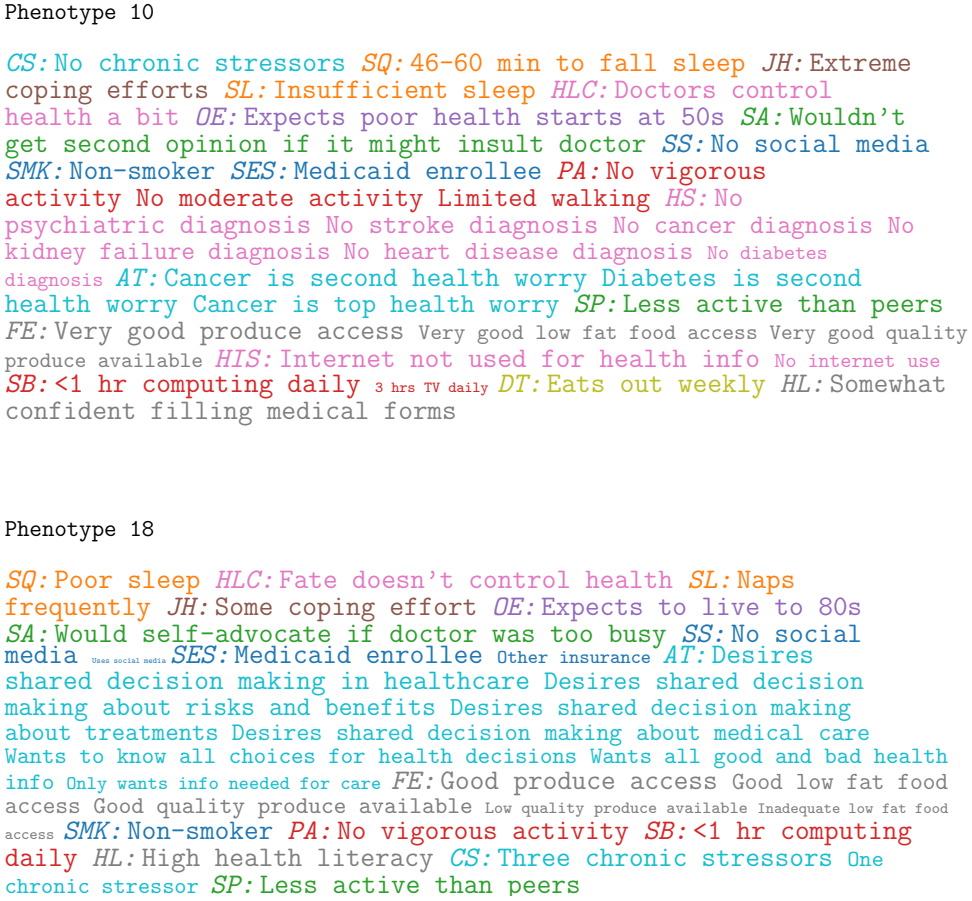

Figure 3.2: Psychosocial Phenotypes Positively Associated with Elevated Weight Status. Phenotypes 10 and 18 were found to be positively associated with elevated weight status. Each word cloud depicts the 20 constructs with lowest normalized entropy for that phenotype and the survey response tokens associated with each construct are shown. The font size is proportional to a token's probability within the given phenotype. The font color is unique to each construct. Only tokens with probability greater than or equal to the phenotype-token prior mean plus 0.5 times the phenotype-token prior standard deviation are shown. Constructs lacking tokens meeting this threshold are omitted. Constructs are visualized in order of increasing normalized entropy. *AC*, alcohol consumption; *PA*, physical activity; *SB*, sedentary behavior; *SL*, sleep; *SMK*, smoking; *DT*, diet; *OE*, outcome expectations; *FE*, food environment; *HL*, health literacy; *PAB*, physical activity barriers; *SS*, social support; *AT*, attitudes; *SP*, self-perception; *HIS*, health information seeking; *SA*, self-advocacy; *HLC*, health locus of control; *CS*, chronic stress; *DM*, demographics; *HS*, health status; *SES*, socioeconomic status; *MH*, mental health; *SQ*, sleep quality; *PS*, perceived stress; *JH*, John Henryism.

To assess between-subject variability, we identified phenotypes above a set threshold for each person in our dataset. Each person had at most three active phenotypes, with most

Figure 3.3: Psychosocial Phenotypes Negatively Associated with Elevated Weight Status. Phenotypes 2, 8, 9, and 15 were negatively associated with elevated weight status. Each word cloud depicts the 20 constructs with lowest normalized entropy for that phenotype and the survey response tokens associated with each construct are shown. The font size is proportional to a token's probability within the given phenotype. The font color is unique to each construct. Only tokens with probability greater than or equal to the phenotype-token prior mean plus 0.5 times the phenotype-token prior standard deviation are shown. Constructs lacking tokens meeting this threshold are omitted. Constructs are visualized in order of increasing normalized entropy. *AC*, alcohol consumption; *PA*, physical activity; *SB*, sedentary behavior; *SL*, sleep; *SMK*, smoking; *DT*, diet; *OE*, outcome expectations; *FE*, food environment; *HL*, health literacy; *PAB*, physical activity barriers; *SS*, social support; *AT*, attitudes; *SP*, self-perception; *HIS*, health information seeking; *SA*, self-advocacy; *HLC*, health locus of control; *CS*, chronic stress; *DM*, demographics; *HS*, health status; *SES*, socioeconomic status; *MH*, mental health; *SQ*, sleep quality; *PS*, perceived stress; *JH*, John Henryism.

people having only one, as illustrated in Figure 3.4 (top). This suggests that phenotypes differentiated between people in most cases. In addition, there was wide variability in how often each phenotype was active across the study population, as illustrated in Figure 3.4 (bottom). This eliminates the possibility that a small number of phenotypes describes all or most of the people in the dataset.

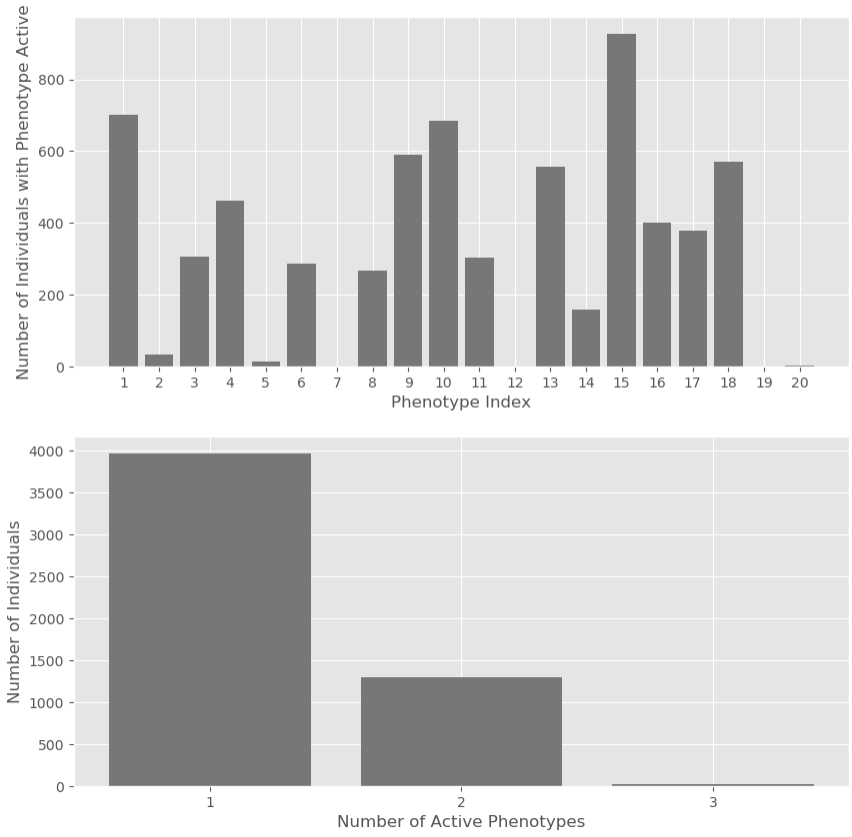

Figure 3.4: Active Phenotypes. For a given individual, a phenotype is considered active if it has probability greater than or equal to the person-phenotype prior probability plus 1 prior standard deviation. Top: Shown is the number of individuals with 1, 2, or 3 active phenotypes; no individuals were found to have 4 or more active phenotypes. Bottom: Shown is the total number of individuals each phenotype is active for.

Discussion

Conceptually meaningful population subgroups The goal of our work was to uncover conceptually meaningful groupings of psychosocial and behavioral characteristics, which could be used to identify and characterize distinct subgroups of individuals within a population. To achieve this goal, we adopted an inductive perspective, using an unsupervised model to find previously unidentified groups of characteristics, or psycho-social phenotypes. This approach constrasts with previous work which takes a deductive, hypothesis-driven approach to claim adverse and favorable "psychosocial profiles of obesity" are unrelated to socioeconomic status (Fuentes, Brondeel, and Franco 2019). Similarly, within the clinical domain, inductive phenotyping methods have been used to move beyond recovering known diagnoses from electronic health records (e.g., Shang, Liu, and Rasmussen 2019) to uncovering previously unknown disease subtypes through modeling patient-generated data (Li et al. 2015; Urteaga, McKillop, and Elhadad 2020). This highlights an inherent tradeoff. MC3M helped us identify novel combinations of psychosocial, behavioral, and contextual factors in the form of psychosocial-behavioral phenotypes. However, as is true for all inductive methods, the model output is completely dependent upon the characteristics of the dataset used for training. Deductive approaches are needed to confirm relationships among intervention strategies, psychosocial-behavioral phenotypes, and behavior targets (Rothman and Sheeran 2020).

Whereas prior work used qualitative methods to discover psychosocial-behavioral phenotypes from in-depth interviews (Burgermaster et al. 2018), the use of MC3M for psychosocial-behavioral phenotyping described here has enabled us to identify psychosocial-behavioral phenotypes in an automated manner. The resulting phenotypes are both novel and plausible. They represent logical and reasonable combinations of psychosocial characteristics, while variations among phenotypes suggest their utility in differentiating meaningful subgroups. Identifying active phenotypes for each individual and assessing the frequency of individuals with an active phenotype across the set of phenotypes (see Figure 3.4) could aid

in prioritizing intervention targets and contextual factors.

Only some of the phenotypes were significantly related to elevated weight status, which suggests the value of approaches that combine psychosocial factors at this level of granularity. Although several phenotypes have features in common, they are present in different combinations. For example, "extreme coping effort" is a survey response token from the John Henryism construct. John Henryism describes the phenomenon of individuals who engage in "high-effort coping" more frequently experiencing hypertension or obesity (Sa 1994; Booth and Cr 2016). Although this phenomenon has been demonstrated to occur among Hispanic Americans (Amw et al. 2015), it was originally identified among Black Americans of low socioeconomic status in the South (Sa 1994). Extreme coping is only present in Phenotypes 10 and 15 and co-occurs with "Medicaid" (a proxy for low-income status) in both cases. However, while Phenotype 10 was positively associated with elevated weight status, Phenotype 15 was not. This could be explained by differences in health literacy between the two phenotypes as John Henryism is characterized by high-effort coping among disadvantaged groups with limited access to tools and assistance. High health literacy could plausibly protect against elevated weight status in Phenotype 15, a hypothesis that could be tested in future work using deductive methods.

Intervention targets in context Behavioral health interventions commonly target modifiable psychosocial and behavioral factors (Glanz, Rimer, and Theory 2015; research 2016). Thus, such interventions may be made more effective if they are designed to take subgroup differences in these factors into account. However, because effect modification is a consistent issue in behavioral interventions (Rothman and Sheeran 2020; Bryan, Tipton, and Ds 2021), the invididual context in which an intervention is applied is also important. Our approach aims to simplify the identification of potential intervention targets while simultaneously presenting them in context. In our case, Phenotype 10 and Phenotype 18 were positively associated with elevated weight status, and also indicated low levels of physical activity

and poor sleep. However, these potential intervention targets should be informed by the contextual factors also present in the phenotypes. For example, both phenotypes highlight low-income status and no computer, internet, or social media use suggesting potential interventions should minimize reliance upon digital media. Meanwhile, concern about chronic disease, good food access, and locus of control oriented to powerful others differentiate Phenotype 10 from Phenotype 18 which is characterized by internal locus of control, high health literacy, and chronic stress. Importantly, the differences in these two phenotypes point to contextual factors on the level of systems and environment that could be better addressed by policy. This underscores both the potential for these types of moderators to influence intervention effects as well as the heterogeneity of psychosocial-behavioral and contextual factors, even within the geographically and ethnically constrained WICER cohort.

Limitations

This work has several limitations. To begin, the dataset used here was obtained from a demographically and geographically constrained cohort. Thus, it is unlikely that the phenotypes we uncovered will generalize to other populations. Nevertheless, our phenotypes did appear to meaningfully cluster individuals within this cohort suggesting that our approach could be useful for tailoring interventions within the community of origin. A second limitation pertains to our use of secondary data. Since we did not collect our data prospectively, we were unable to select the constructs to be measured. Though we applied out method to data representing a relatively large set of psychosocial, behavioral and social determinants of health, these did not comprehensively represent all psychosocial-behavioral constructs of interest. We would expect that expanding the dataset to incorporate such constructs would result in learning a different set of phenotypes.

3.2 Aim 1B. *Identify MC3M phenotypes associated with a health outcome.*

Research Question

Can MC3M phenotypes be used to predict the presence of a targeted health outcome?

Methods

Predictive Modeling. To determine which MC3M phenotypes were associated with elevated weight status, we first use MC3M's person-phenotype distributions as covariate vectors in a binary prediction model targeting elevated weight status. We then rely on significance testing to identify which phenotypes were positively and negatively associated with elevated weight status.

For our predictive model we use logistic regression with an elastic net (EN) penalty (Zou and T. 2005). We tune the EN logistic regression hyperparameters with 10-fold cross-validation. We then use the best performing hyperparameters to train a model on the full training set, and significant odds ratios are used to determine which phenotypes are positively or negatively associated with elevated weight status. Because the values of our covariates are constrained to the interval [0, 1], we first apply a log transformation to the person-phenotype distributions. This removes the constraint and results in more stable training.

EN logistic regression has several hyperparameters including the regularization strength, λ, and the L1/L2 elastic net mixing parameter, γ. We optimize λ and γ using 10-fold cross validation and grid search: $\lambda \in \{0.01, 0.5, 0.1, 1.0, 5.0, 10.0\}$, $\gamma \in \{0.0, 0.1, 0.2, \ldots, 0.9, 1.0\}$. The areas under the ROC (AUROC) and Precision-Recall (AUPRC) curves were used as target metrics. For each hyperparameter setting, we generate a point estimate for each metric and bootstrapped 95% confidence intervals (CIs). Bootstrapped metric distributions are generated by pooling true labels and label predictions over all held-out folds, sampling from this pool with replacement, and calculating the metric for each sample (10,000 samples

for each hyperparameter setting).

We use the following bootstrapping procedure to identify significant regression coefficients. First, we sample instances with replacement from the training set such that the number of positive and negative cases is preserved. Next, we train an EN logistic regression model, and recorded the fitted coefficient values. This process is repeated 10,000 times yielding bootstrap distributions for each coefficient. We used these distributions to define bootstrapped 95% CIs. Significant coefficients were taken as those whose 95% CIs were non-overlapping with the null value, 0.

Results

Psychosocial-Behavioral Phenotypes and Weight Status Figure 3.5 shows relationships between phenotypes and weight status according to our EN logistic regression. The bootstrap 95% CIs for the model odds ratios from our best performing EN logistic regression indicated that six phenotypes – 2, 8, 9, 10, 15, and 18 – had coefficients significantly different from the null. Two (Phenotypes 10 and 18) were significantly positively associated with elevated weight status while the remaining four (Phenotypes 2, 8, 9, and 15) were significantly negatively associated with weight status.

Phenotype 10 and Phenotype 18 were positively associated with elevated weight status. These phenotypes are described here and presented in Figure 3.2. Phenotype 10 (n=695, 13%) was positively associated with elevated weight status (OR= 1.02; 95% CI = 1.01, 1.03). This phenotype was characterized by a perception of low stress along with extreme coping in the face of discrimination (i.e., John Henryism), poor sleep, little physical activity, low income, worry about chronic disease, and being unlikely to self-advocate within the healthcare system. Phenotype 18 (n=588, 11%) was also positively associated with elevated weight status (OR= 1.03; 95% CI = 1.02, 1.04). This phenotype was characterized by poor sleep, little physical activity, low income, good food access, high health literacy, willingness to self-advocate and interest in shared decision making within the healthcare system, and

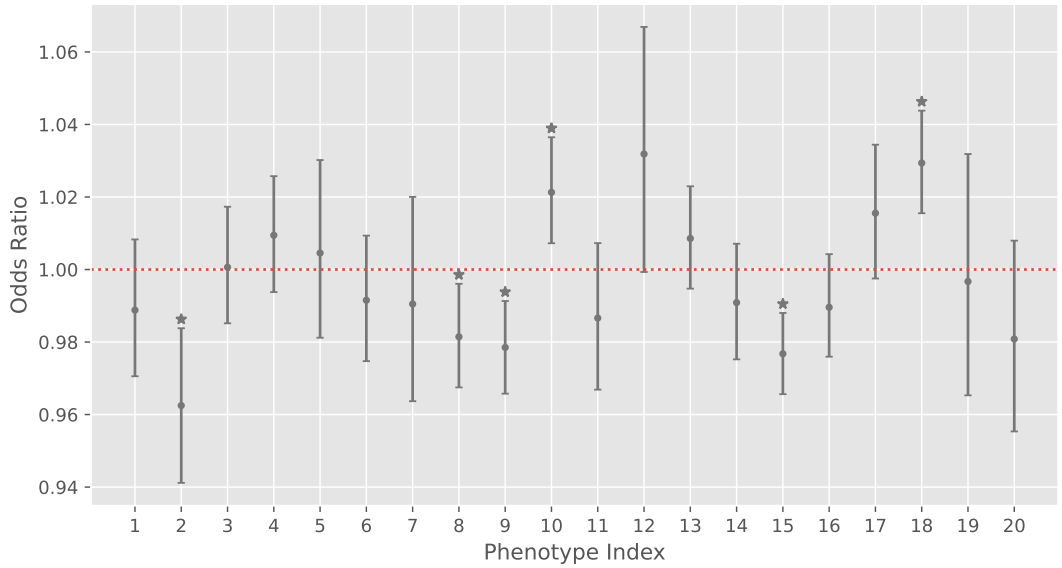

Figure 3.5: Odds Ratios for Elevated Weight Status Predictive Model. Points and error bars represent bootstrap means and 95% CIs, respectively, for EN logistic regression odds ratios. Log-transformed person-phenotype distributions were used as covariates. Elevated weight status was used as the predictive target. Stars indicate coefficients odds ratios whose 95% CIs do not contain the null.

high stress.

Four psychosocial-behavioral phenotypes were negatively associated with elevated weight status; they are described here and presented in Figure 3.3. Phenotype 15 (n=946, 18%) was negatively associated with elevated weight status (OR= 0.98 95% CI = 0.97, 0.99). This phenotype was characterized by poor sleep; no chronic stressors, but extreme coping in the face of discrimination; Hispanic immigrant to the US with more than two adults living in their household; self-perception as active, but no vigorous activity reported; no chronic disease diagnoses, very high life expectancy; good food access; and low income. Phenotype 9 (n=601, 11%) was negatively associated with elevated weight status (OR=0.98; 95% CI=0.97, 0.99). This phenotype was characterized by poor sleep, internal health locus of control; willingness to self-advocate and interest in shared decision making within the healthcare system; low stress; no barriers to physical activity, self-perception as comparably active to peers, walking but no vigorous activity reported; social media use; low income; high

health literacy; and good food access. Phenotype 8 (n=294, 6%) was negatively associated with elevated weight status (OR=0.98; 95% CI0.97, 0.99). This phenotype was characterized by alcohol consumption; self-perception as more active than peers; positive outlook on health, no chronic disease diagnoses, very high internal locus of control in health and low willingness to self-advocate within the healthcare system; more than two adults living in their household; low income; very good food access; and no mental health issues. Phenotype 2 (n=50, 1%) was negatively associated with elevated weight status (OR=0.96; 95% CI=0.94, 0.98). This phenotype was characterized by alcohol consumption; self-perception as more active than peers, but little physical activity reported; very good food access; good sleep quality; low expectations for a long, healthy life; no mental health issues; computer and social media use, though not for health information seeking; and working for pay.

Discussion

Linking Phenotypes to Weight Status Our approach yielded 20 phenotypes, of which six were positively or negatively related to weight status. Each phenotype includes features that have previously been associated with risk for or protection from elevated weight status. For example, multiple chronic stressors are present only in Phenotype 18, which was positively related to elevated weight status. Chronic stress has been directly related to obesity among Latinas (Stanhope, Picon, and Schlusser 2021), and a mechanistic pathway relating stress to elevated weight status has been established (Xiao et al. 2020). This highlights the potential value of psychosocial-behavioral phenotyping for understanding within-group heterogeneity. We note that the magnitude of associations between phenotypes and elevated weight status, though significant, are relatively small. Nevertheless, they remain potentially meaningful in the context of population health interventions (Matthay, Hagan, and Gottlieb 2021). Conditional upon confirmatory deductive analysis, weight-loss interventions targeting phenotypes positively associated with elevated weight status could result in health benefits on a sub-population level.

Limitations

This work had several limitations. First, BMI is at best a proxy measure for individual health. We rely on BMI to assess weight status, primarily due to convenience; weight and height measurements are available for everyone in the sample population. Though linking psychosocial-behavioral phenotypes to BMI-derived weight status is convenient and helpful for demonstrating the potential utility of psychosocial-behavioral phenotypes in designing interventions, we would not encourage others to rely on BMI in this way.

Chapter 4

Aim 2. Develop a method for learning phenotypes from partially labeled clinical data.

Phenotypes are powerful tools for working with observational clinical data in the absence of reliable disease labels (Hripcsak and Albers 2012). Disease-specific phentoypes allow researchers to sift through large-scale clinical data stores to identify patients with evidence of specific clinical conditions. By answering the question of who has what disease, phenotypes power essential tasks such as cohort selection, trial recruitment and clinical outcome prediction (Hripcsak and Albers 2012; Richesson et al. 2013; Richesson et al. 2016).

Traditionally, phenotypes were developed by groups of clinical experts who painstakingly hand-tuned rule-based algorithms. The limited scalability of this approach has led to the development of automated methods for learning phenotypes directly from clinical data. Many studies in this vein utilize supervised machine learning methods to build phenotyping algorithms (Bergquist et al. 2017; Esteban et al. 2017b). Though this approach avoids laborious expert knowledge engineering, it requires significant amounts of labeled clinical data generated by manual chart review.

To avoid costly, expert-generated disease labels, many authors have utilized unsupervised methods to cluster patients according to underlying patterns in their clincal data (Joshi et al. 2016; Ho et al. 2014b; Ho, Ghosh, and Sun 2014b; Wang et al. 2015b; Miotto et al. 2016b)i. In this setting, such patterns play the role of phenotypes. Unsupervised phenotyping methods often learn multiple phenotypes simultaneouly, which may confer evidence of specific diseases. However, such phenotypes are generally not guaranteed to represent single disease concepts.

This complicates their evaluation and use in downstream tasks.

In this aim, we propose the Semi-Supervised Mixed Membership Model (SS3M), a probabilistic graphical model which utilizes relatively few disease labels to learn multiple disease-specific phenotypes from multi-modal observational clinical data. SS3M addresses the limitations of supervised phenotyping by reducing the amount of labeled data needed to learn disease phenotypes; disease labels are not required for all patients, and labeled patients need not possess labels for all diseases. SS3M also addresses the limitations of unsupervised phenotyping by associating disease labels with the phenotypes to be learned; a label specifies which disease a phenotype is meant to represent. This simplifies both the qualitative and quantitative evaluation of SS3M phenotypes. Qualitatively, phenotype labels inform us as to what content we should expect to be well represented within a learned phenotype. Quantitatively, we can evaluate how well learned phenotypes predict labels on a held-out patient cohort using standard performance metrics.

4.1 Aim 2A. *Derive and implement the Semi-Supervised Mixed Membership Model (SS3M) — a probabilistic graphical model for learning interpretable, disease-specific phenotypes from partially labeled, multi-modal clinical data.*

Background

Disease phenotyping can often be reduced to classification. However, training supervised phenotyping models with electronic health records data is challenging due to the lack of gold standard labels. Though fully labeling a clinical dataset is generally infeasible, obtaining a small amount labeled data may be possible. Here we develop Semi-Supervised Mixed Membership Models (SS3M) — a family of semi-supervised models for learning phenotypes from partially labeled data. We build SS3M by incorporating a novel semi-supervision mechanism

into an otherwise fully unsupervised model previously proposed for learning interpretable phenotypes from heterogenous clinical data. We derive and implement a Markov Chain Monte Carlo sampler to perform posterior inference on the model's latent variables. We then evaluate SS3M in simulation. First we simulate data from SS3M's generative model parameterized with a set of ground truth phenotypes and their corresponding labels. We then fit these data with a randomly initialized SS3M model and check to see if the model can recover the ground truth phenotypes in both structure and identity.

Research Question

Can SS3M recover ground-truth phenotypes from partially labeled data?

Methods

Model. Here we provide a detailed description of SS3M's structure. The model's generative process and graphical model provide complementary perspectives and are detailed in Algorithm 1 and Figure 4.1 respectively. Table 4.1 provides descriptions of all model variables. Here and in the rest of the text, we use bold capital letters to indicate groups of variables and indices to refer to subsets or specific elements. A bold capital letter without indices indicates all variables within the group. A colon within a variable subscript indicates all elements within the corresponding dimension.

Let D, S, and P be the number of patients, clinical data sources and phenotypes, respectively. Each patient $d \in \{1, \ldots, D\}$ is associated with several sets of tokenized clinical observations $\boldsymbol{W}_{sd:}$ (e.g. medication names), one for each data source $s \in \{1, .., S\}$. In addition, each patient has a set of partially observed binary labels. Patient d's labels specify the values of their phenotype activations, $\boldsymbol{A}_{d:}$, thereby indicating for them which phenotypes $p \in \{1, \ldots, P\}$ are set to be "on" or "off". A latent phenotype assignment, $\boldsymbol{Z}_{sdn}$, is assigned to each observation, $\boldsymbol{W}_{sdn}$. Each assignment is drawn from a categorical patient-phenotype distribution parameterized by a normalized P-dimensional vector, $\boldsymbol{\Theta}_d$. A phenotype assign-

ment specifies which phenotype-token distribution an observation was drawn from. Each phenotype-token distribution is parameterized by a normalized V_s-dimensional vector, $\boldsymbol{\Phi}_{sp}$, where V_s is the size of the vocabulary for data source s.

Algorithm 1 Generative process for SS3M

Initialize: α, β, β^*, $\{\gamma_s\}_{s=1}^{S}$

 # Sample global variables
Sample $B^* \sim \text{Gamma}(\beta^*)$
for *each phenotype* $p = 1$ **to** P **do**
  Sample $\boldsymbol{B}_p \sim \text{Gamma}(\beta)$
  Sample $\boldsymbol{C}_p \sim \text{Beta}(\alpha)$
  for *each data source* $s = 1$ **to** S **do**
    Sample $\boldsymbol{\Phi}_{sp} \sim \text{Dirichlet}(\gamma_s)$
  end
end
 # Sample local variables
for *each patient* $d = 1$ **to** D **do**
  for *each phenotype* $p = 1$ **to** P **do**
    Sample $\boldsymbol{A}_{dp} \sim \text{Bernoulli}(\boldsymbol{C}_p)$
  end
  Sample $\boldsymbol{\Theta}_d \sim \text{Dirichlet}(\boldsymbol{A}_{d:} \odot \boldsymbol{B}_{:} + (\mathbb{1} - \boldsymbol{A}_{d:})B^*)$
  for *each data source* $s = 1$ **to** S **do**
    for *each observation* $n = 1$ **to** N_{sd} **do**
      Sample $\boldsymbol{Z}_{sdn} \sim \text{Categorical}(\boldsymbol{\Theta}_d)$
      Sample $\boldsymbol{W}_{sdn} \sim \text{Categorical}(\boldsymbol{\Phi}_{s\boldsymbol{Z}_{sdn}})$
    end
  end
end

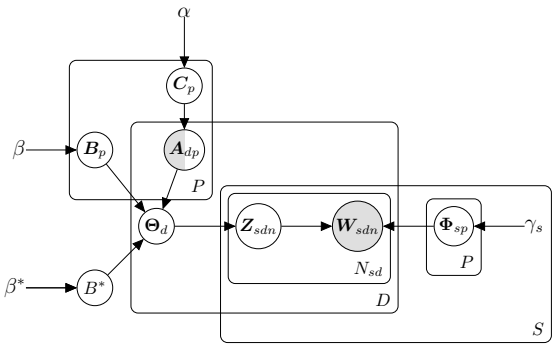

Figure 4.1: Graphical model for SS3M

Variable	Description
D	Number of patients
S	Number of data sources
P	Number of phenotypes
N_{sd}	Number of tokens of source s for patient d
V_s	Size of vocabulary for source s
$\boldsymbol{A}$	Phenotype activations (partially observed)
$\boldsymbol{B}$	Active phenotype parameters
B^*	Inactivate phenotype parameter
$\boldsymbol{C}$	Phenotype prevalences
$\boldsymbol{\Theta}$	Patient-phenotype distribution parameters
$\boldsymbol{\Phi}$	Phenotype-token distribution parameters
$\boldsymbol{Z}$	Phenotype assignments
$\boldsymbol{W}$	Token observations
α	Hyperparameters
β	
β^*	
γ	

Table 4.1: SS3M variable descriptions

A patient's label set directly impacts their patient-phenotype distribution, and thereby all their assignments. This is due to the roles of $\boldsymbol{A}$, $\boldsymbol{B}$ and B^* in parameterizing the Dirichlet distributions on the elements of $\boldsymbol{\Theta}$:

$$\boldsymbol{\Theta}_d \sim \text{Dirichlet}(\boldsymbol{A}_{d:} \odot \boldsymbol{B}_{:} + (\mathbb{1} - \boldsymbol{A}_{d:})B^*)$$

where $\odot$ indicates element-wise multiplication. When $\boldsymbol{A}_{dp} = 1$, patient d has phenotype p "on"; $\boldsymbol{B}_p$ is used to parameterize the p^{th} dimension of the Dirichlet on $\boldsymbol{\Theta}_d$. When $\boldsymbol{A}_{dp} = 0$, the phenotype is "off", and B^* is used instead. The hyperparameters β and β^* parameterize the gamma distributions on $\boldsymbol{B}$ and B^* such that the model is encouraged to sample values of $\boldsymbol{B}_p$ and B^* to maintain $\boldsymbol{B}_p > 1$ and $B^* < 1$. In this setting, when $\boldsymbol{A}_{dp} = 1$ the values of $\boldsymbol{\Theta}_d$ push the patient-phenotype distribution toward allocating more probablity mass for phenotype p. This in turn, results in a larger proportion of patient d's observations being assigned to phenotype p. During inference, this mechansim forms the connection between labels, activations and the content of phenotypes. Labels set phenotype activations "on" or "off" for each patient. For each patient, phenotypes that are "on" account for the majority of phenotype assignments. Thus, labels, by way of activations, strongly influence the quantity of observations that are funneled toward learning any given phenotype.

Activations are *partially observed*. If a patient d has an observed binary label for phenotype p, then the value of $\boldsymbol{A}_{dp}$ is held fixed at the observed value. If the label is unobserved, then the model samples the value of $\boldsymbol{A}_{dp}$ during inference. In this latter case, $\boldsymbol{A}_{dp}$ is modeled as a binary variable drawn from a Bernoulli distribution parameterized by $\boldsymbol{C}_p$ — a beta-distributed latent variable controlling the likelihood of phenotype p being "on" within the patient population (i.e. $\boldsymbol{C}_p$ estimates the prevalence of phenotype p). This handling of partially observed labels is what allows SS3M to function as a semi-supervised model.

SS3M can handle both semi-supervised phenotypes for which we have some number of labels, as well as unsupervised phenotypes that lack labels all together. This is a useful property when applying the model to clinical data. In this setting we are unlikely to have labels for all the conditions represented in our dataset. The structure of the conditions we lack labels for can be targeted by SS3M's unsupervised phenotypes during inference. Since the set of phenotypes underlying the data need not be limited to the labeled set, including unsupervised phenotypes can help semi-supervised phenotypes "focus" on capturing those phenotypes which best align with their labels.

Inference. We implement a collapsed Gibbs sampler to obtain posterior estimates of our model's latent variables. The variables $\boldsymbol{C}$, $\boldsymbol{\Theta}$, and $\boldsymbol{\Phi}$ are easily integrated out of the joint distribution due to conjugate relationships between their distributions and those on $\boldsymbol{A}$, $\boldsymbol{Z}$, and $\boldsymbol{W}$, respectively. The collapsed joint's complete conditional distributions for the elements of $\boldsymbol{A}$ and $\boldsymbol{Z}$ are discrete, easily normalized, and can be sampled from directly. However, the complete conditionals for $\boldsymbol{B}$ and B^* do not have closed forms. We use Hamiltonian Monte Carlo to sample from these (Neal 2011). We set our path length to $L = 15$ and step size to $\epsilon = 10^{-3}$, as these parameters yielded stable trajectories with high acceptance rates in preliminary experiments.

Below we derive the necessary components for our collapsed Gibbs samples: the collapsed joint distribution, complete conditionals, and the potential gradients used in HMC.

Collapsed Joint. The joint distribution for SS3M is

$$p(\boldsymbol{A}, \boldsymbol{B}, B^*, \boldsymbol{C}, \boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{Z}, \boldsymbol{W}; \alpha, \beta, \beta^*, \gamma) = p(B^*; \beta^*) \prod_{p=1}^{P} p(\boldsymbol{B}_p; \beta) \tag{4.1}$$

$$\times\, p(\boldsymbol{C}_p; \alpha) \prod_{d=1}^{D} p(\boldsymbol{\Theta}_d | \boldsymbol{A}_{d:}, \boldsymbol{B}, B^*) p(\boldsymbol{A}_{dp} | \boldsymbol{C}_p) \prod_{s=1}^{S} p(\boldsymbol{\Phi}_{sp}; \gamma_s) \prod_{n=1}^{N_{sd}} \tag{4.2}$$

$$\times\, p(\boldsymbol{W}_{sdn} | \boldsymbol{Z}_{sdn}, \boldsymbol{\Phi}_{s:}) p(\boldsymbol{Z}_{sdn} | \boldsymbol{\Theta}_d).$$

The distribution for each factor on the RHS is given in the generative process described in Algorithm 1. We integrate $\boldsymbol{C}$,$\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ out of SS3M's joint distribution to obtain the

collapsed joint:

$$p(\boldsymbol{A}, \boldsymbol{B}, B^*, \boldsymbol{Z}, \boldsymbol{W}; \alpha, \beta, \beta^*, \gamma) \tag{4.3}$$

$$= p(B^*; \beta^*) \prod_{p=1}^{P} p(\boldsymbol{B}_p; \beta) \prod_{s=1}^{S} \prod_{d=1}^{D} \prod_{n=1}^{N_{sd}} \int_{\boldsymbol{C}_p} p(\boldsymbol{A}_{dp}|\boldsymbol{C}_p) p(\boldsymbol{C}_p; \alpha) d\boldsymbol{C}_p \tag{4.4}$$

$$\times \int_{\boldsymbol{\Theta}_d} p(\boldsymbol{Z}_{sdn}|\boldsymbol{\Theta}_d) p(\boldsymbol{\Theta}_d|\boldsymbol{A}_{d:}, \boldsymbol{B}, B^*) d\boldsymbol{\Theta}_d \int_{\boldsymbol{\Phi}_{sp}} p(\boldsymbol{\Phi}_{sp}; \gamma_s) p(\boldsymbol{W}_{sdn}|\boldsymbol{Z}_{sdn}, \boldsymbol{\Phi}_{s:}) d\boldsymbol{\Phi}_{sp}$$

$$= p(B^*; \beta^*) \prod_{p=1}^{P} p(\boldsymbol{B}_p; \beta) \prod_{d=1}^{D} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{\Gamma(\alpha_1 + \sum_d \boldsymbol{A}_{dp}) \Gamma(\alpha_2 + D - \sum_d \boldsymbol{A}_{dp})}{\Gamma(\alpha_1 + \alpha_2 + D)} \tag{4.5}$$

$$\times \frac{\Gamma(\sum_p \boldsymbol{r}_{dp})}{\prod_p \Gamma(\boldsymbol{r}_{dp})} \frac{\prod_p \Gamma(\boldsymbol{r}_{dp} + n_{dp})}{\Gamma(\sum_p \boldsymbol{r}_{dp} + n_{dp})} \prod_{s=1}^{S} \frac{\Gamma(\sum_v \gamma_{sv})}{\prod_v \Gamma(\gamma_{sv})} \frac{\prod_v \Gamma(\gamma_{sv} + n_{spv})}{\Gamma(\sum_v \gamma_{sv} + n_{spv})},$$

where $\Gamma(\cdot)$ indicates the Gamma function, $\boldsymbol{r}_{dp} = \boldsymbol{A}_{dp} \odot \boldsymbol{B}_p + (\mathbb{1} - \boldsymbol{A}_{dp}) B^*$, n_{dp} is the number of patient d's observations assigned to phenotype p, and n_{spv} is the number of times token v from data source s has been assigned to phenotype p.

Complete Conditionals. Here we obtain proportionalities for the complete conditional distributions of each latent variable in our collapsed joint. Note we use "$-$" to indicate all variables in the joint *excluding* that which appears on the left side of the conditioning bar.

$$p(\boldsymbol{A}_{dp}|-) \propto \Gamma(\alpha_1 + \sum_{d'} \boldsymbol{A}_{d'p}) \Gamma(\alpha_1 + D - \sum_{d'} \boldsymbol{A}_{d'p}) \frac{\Gamma(\sum_{p'} \boldsymbol{r}_{dp'})}{\Gamma(\boldsymbol{r}_{dp})} \frac{\Gamma(\boldsymbol{r}_{dp} + n_{dp})}{\Gamma(\sum_{p'} \boldsymbol{r}_{dp'} + n_{dp'})} \tag{4.6}$$

$$p(\boldsymbol{B}_p|-) \propto \text{Gamma}(\boldsymbol{B}_p; \beta) \prod_{d=1}^{D} \frac{\Gamma(\sum_{p'} \boldsymbol{r}_{dp'})}{\Gamma(\boldsymbol{r}_{dp})} \frac{\Gamma(\boldsymbol{r}_{dp} + n_{dp})}{\Gamma(\sum_{p'} \boldsymbol{r}_{dp'} + n_{dp'})} \tag{4.7}$$

$$p(B^*|-) \propto \text{Gamma}(B^*; \beta^*) \prod_{d=1}^{D} \frac{\Gamma(\sum_p \boldsymbol{r}_{dp})}{\prod_p \Gamma(\boldsymbol{r}_{dp})} \frac{\prod_p \Gamma(\boldsymbol{r}_{dp} + n_{dp})}{\Gamma(\sum_p \boldsymbol{r}_{dp} + n_{dp})} \tag{4.8}$$

$$p(\boldsymbol{Z}_{sdn}|-) \propto (\boldsymbol{r}_{dp} + n_{dp}^{-sdn}) \frac{\gamma_{sv} + n_{spv}^{-sdn}}{\sum_{v'} \gamma_{sv'} + n_{spv'}^{-sdn}} \tag{4.9}$$

In the proportionality for $p(\boldsymbol{Z}_{sdn}|-)$, the $n_{\cdot}^{-sdn}$ terms indicate total token assignment counts excluding the current assignment, $\boldsymbol{Z}_{sdn}$. The index v refers to the observed value of $\boldsymbol{W}_{sdn}$. The proportionalities for $p(\boldsymbol{A}_{dp}|-)$ and $p(\boldsymbol{Z}_{sdn}|-)$ are simple to normalize, and can be sam-

pled from directly afterward. This is not the case for $p(\boldsymbol{B}_p|-)$ and $p(B^*|-)$, which we sample from using Hamiltonian Monte Carlo (HMC).

Hamiltonian Monte Carlo. To use HMC we must calculate a potential energy function proportional to our target distribution and calculate its gradient with respect to the corresponding random variable. Note that the $\boldsymbol{B}$ and B^* are constrained to $\mathbb{R}^+$. We remove this constraint by applying a change of variables to sample in log space.

$$p(\hat{\boldsymbol{B}}_p|-) \propto \exp(\hat{\boldsymbol{B}}_p\beta_1 - \exp(\hat{\boldsymbol{B}}_p)/\beta_2) \prod_{d=1}^{D} \frac{\Gamma(\sum_{p'} \hat{\boldsymbol{r}}_{dp'})}{\Gamma(\hat{\boldsymbol{r}}_{dp})} \frac{\Gamma(\hat{\boldsymbol{r}}_{dp} + n_{dp})}{\Gamma(\sum_{p'} \hat{\boldsymbol{r}}_{dp'} + n_{dp'})} \tag{4.10}$$

$$p(\hat{B}^*|-) \propto \exp(\hat{B}^*\beta_1^* - \exp(\hat{B}^*)/\beta_1^*) \prod_{d=1}^{D} \frac{\Gamma(\sum_{p} \hat{\boldsymbol{r}}^*_{dp})}{\prod_p \Gamma(\hat{\boldsymbol{r}}^*_{dp})} \frac{\prod_p \Gamma(\hat{\boldsymbol{r}}^*_{dp} + n_{dp})}{\Gamma(\sum_{p} \hat{\boldsymbol{r}}^*_{dp} + n_{dp})}, \tag{4.11}$$

where $\hat{\boldsymbol{B}}_p = \log \boldsymbol{B}_p$, $\hat{B}^* = \log B^*$, $\hat{\boldsymbol{r}}_{dp} = \boldsymbol{A}_{dp} \odot \exp(\hat{\boldsymbol{B}}_p) + (\mathbb{1} - \boldsymbol{A}_{dp})B^*$, and $\hat{\boldsymbol{r}}^*_{dp} = \boldsymbol{A}_{dp} \odot \boldsymbol{B}_p + (\mathbb{1} - \boldsymbol{A}_{dp}) \exp(\hat{B}^*)$.

The potentials we require are obtained by negating the log of our transformed target distributions.

$$U(\hat{\boldsymbol{B}}_p) = -\log p(\hat{\boldsymbol{B}}_p|-) \tag{4.12}$$

$$\propto \exp(\hat{\boldsymbol{B}}_p)/\beta_2 - \hat{\boldsymbol{B}}_p\beta_1 \tag{4.13}$$

$$- \sum_{d=1}^{D} \log\Gamma(\sum_{p'} \hat{\boldsymbol{r}}_{dp'}) - \log\Gamma(\hat{\boldsymbol{r}}_{dp}) + \log\Gamma(\hat{\boldsymbol{r}}_{dp} + n_{dp}) - \log\Gamma(\sum_{p'} \hat{\boldsymbol{r}}_{dp'} + n_{dp'})$$

$$U(\hat{B}^*) = -\log p(\hat{B}^*|-) \tag{4.14}$$

$$\propto \exp(\hat{B}^*)/\beta_1^* - \hat{B}^*\beta_1^* \tag{4.15}$$

$$- \sum_{d=1}^{D} \log\Gamma(\sum_{p} \hat{\boldsymbol{r}}^*_{dp}) - \log\Gamma(\sum_{p} \hat{\boldsymbol{r}}^*_{dp} + n_{dp}) + \sum_{d=1}^{D} \sum_{p=1}^{P} \log\Gamma(\hat{\boldsymbol{r}}^*_{dp}) - \log\Gamma(\hat{\boldsymbol{r}}^*_{dp} + n_{dp})$$

Their gradients are as follows.

$$\frac{\partial U(\hat{\boldsymbol{B}}_p)}{\partial \hat{\boldsymbol{B}}_p} = -\beta_1 + \exp(\hat{\boldsymbol{B}}_p)[\frac{1}{\beta_2} + \sum_{d=1}^{D}\{\Psi(\hat{\boldsymbol{r}}_{dp}) - \Psi(\sum_{p'}\hat{\boldsymbol{r}}_{dp'}) + \Psi(\sum_{p'}\hat{\boldsymbol{r}}_{dp'} + n_{dp'}) - \Psi(\hat{\boldsymbol{r}}_{dp} + n_{dp})\}\boldsymbol{A}_{dp}] \tag{4.16}$$

$$\frac{\partial U\left(\hat{B}^*\right)}{\partial \hat{B}^*} = -\beta_1^* + \exp\left(\hat{B}^*\right)\left[\frac{1}{\beta_2^*} + \sum_{d=1}^{D}\sum_{p=1}^{P}\left\{\Psi\left(\hat{\boldsymbol{r}}_{dp}^*\right) - \Psi\left(\sum_{p'}\hat{\boldsymbol{r}}_{dp'}^*\right) + \Psi\left(\sum_{p'}\hat{\boldsymbol{r}}_{dp'}^* + n_{dp'}\right) - \Psi\left(\hat{\boldsymbol{r}}_{dp}^* + n_{dp}\right)\right\}(1 - \boldsymbol{A}_{dp})\right], \tag{4.17}$$

where $\Psi(\cdot)$ indicates the Digamma function.

As detailed in Neal 2011, given a step size, ϵ, and path length, L, these gradients allow us to integrate trajectories in log space to arrive at new candidate states for our random variables. We then evaluate the total energy change using our potential energy functions to decide whether to accept or reject our candidate states.

Collapsed Gibbs Sampler. We now have all the necessary elements to construct a collapsed Gibbs sampler for SS3M. The procedure is described below in Algorithm 2.

Algorithm 2 Collapsed Gibbs Sampler for SS3M

Intialize: α, β, β^*, $\{\gamma_s\}_{s=1}^S$
Sample: $\boldsymbol{A}, \boldsymbol{B}, B^*, \boldsymbol{Z}$ from their priors in the complete joint
Load: tokenized observations into $\boldsymbol{W}$ & labels into $\boldsymbol{A}$

for *each iteration* **do**
  for *each patient* $d \in \{1, \ldots, D\}$ **do**
    for *each data source* $s \in \{1, \ldots, S\}$ **do**
      for *each observation* $n \in \{1, \ldots, N_{sd}\}$ **do**
        Sample $\boldsymbol{Z}_{sdn} \sim p(\boldsymbol{Z}_{sdn}|-)$
      end
    end
    for *each phenotype* $p \in \{1, \ldots, P\}$ **do**
      if $\boldsymbol{A}_{dp}$ *does not have a fixed label* **then**
        Sample $\boldsymbol{A}_{dp} \sim p(\boldsymbol{A}_{dp}|-)$
      end
    end
  end
  for *each phenotype* $p \in \{1, \ldots, P\}$ **do**
    $\boldsymbol{B}_p \leftarrow \exp(HMC(\hat{\boldsymbol{B}}_p, U(\hat{\boldsymbol{B}}_p), \nabla_{\hat{\boldsymbol{B}}_p} U(\hat{\boldsymbol{B}}_p), \epsilon, L))$
  end
  $B^* \leftarrow \exp(HMC(\hat{B}^*, U(\hat{B}^*), \nabla_{\hat{B}^*} U(\hat{B}^*), \epsilon, L))$
end

Return: Samples of $\boldsymbol{A}, \boldsymbol{B}, B^*, \boldsymbol{Z}$

Data. We create simulated patient cohorts by using ancestral sampling to draw observations and labels from our model. To begin, we define 10 ground truth phenotypes, $\boldsymbol{\Phi}_{\text{true}}$, in a manner inspired by Griffiths and Steyvers 2004. Each phenotype is a set of three categorical distributions defined over three separate vocabularies each of length 25. This allows us to visualize each phenotype as a set of three 5×5 grids. Each component of a phenotype places uniform probability mass over 5 tokens corresponding to a row or column in the grid. Next, we simulate ground truth labels for each patient. This is done by first drawing values for $\boldsymbol{C}$ from Beta distributions parameterized with $\gamma = (10^2, 10^3)$. This ensures each phenotype is active in about 10% of the cohort. We then use the values of $\boldsymbol{C}$ to draw an array of ground truth labels, $\boldsymbol{A}_{\text{true}}$. The values of $\boldsymbol{B}$ and B^* are set to $(10., \ldots, 10.)$ and 10^{-2}, respectively. These values ensure observations are highly likely to be drawn from active phenotypes. Fi-

nally, we draw values for $\boldsymbol{\Theta}$ and $\boldsymbol{Z}$ which, along with $\boldsymbol{\Phi}_{\text{true}}$, are used to generate our simulated observations, $\boldsymbol{W}$.

Evaluation. We expose SS3M to simulated cohort data and run our inference algorithm to recover the ground truth phenotypes. Our evaluation is qualitative in nature. We check to confrim that (1) the SS3M's phenotypes are similar in structure to the ground truth phenotypes and (2) the identities of each phenotype (i.e. the label index values) match the ground phenotype identities.

We use the same training set of observations for each of our experiments. For each experiment, we produce a label set by downsampling ground truth labels in $\boldsymbol{A}_{\text{true}}$. We run 2 experiments in which we retain 1% and 5% of positive labels. We then sample negative labels to match the total number of positive lables for a given phenotype.

Results & Discussion

Figure 4.2 shows the results of our simulated studies. When training on 1% of available labels, SS3M struggles to recover ground truth. Some of the inferred phenotypes appear to be superpositions of multiple ground truth phenotypes. Though some of the phenotype-token distributions do indeed mirror ground truth, many of the indices are mismatched. Full recovery of ground truth requires both the recovery of phenotype structure as well as phenotype identity. Both of these requirements are met when SS3M is exposed to just 5% of available labels. For our dataset of 1000 simulated patients, 5% of labels corresponds to 14-15 labeled patients per phenotype – half labeled positive and half labeled negative.

Limitations

The inference algorithm we have developed for SS3M relies on MCMC, which can take many iterations to converge, particularly as the number of observations grows large. This limits the applicability of SS3M to very large clinical data sets where more efficient algorithms

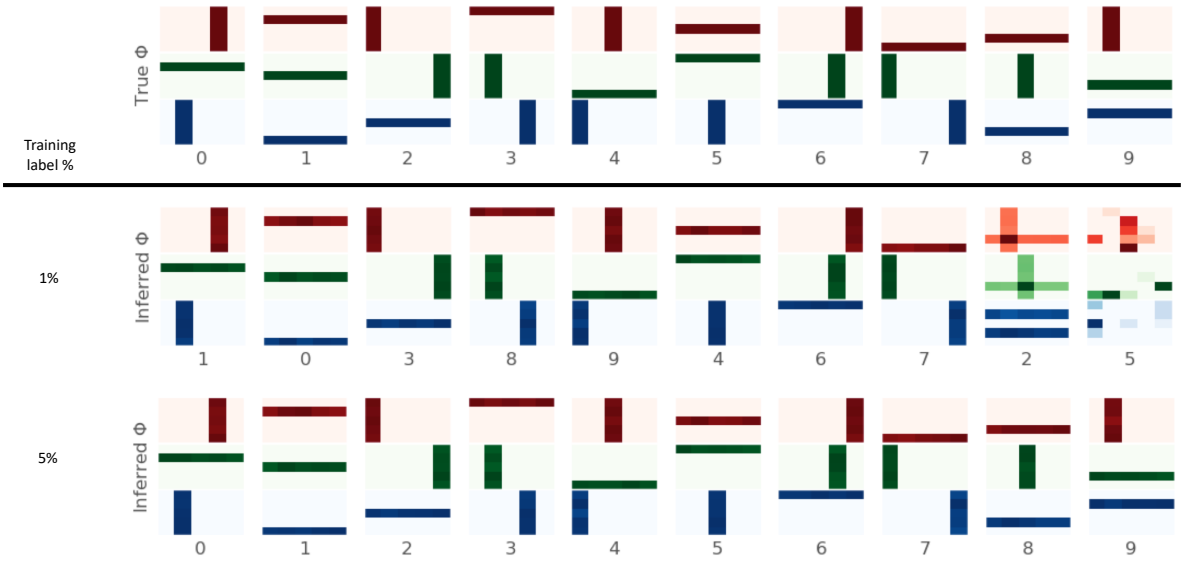

Figure 4.2: Phenotype inference for simulated patient cohort. **Top row**: Ground truth phenotypes. **Middle & Bottom rows**: Phenotypes inferred using 1% and 5% of ground truth labels for training. When training on 5% of available labels, SS3M recovers ground truth phenotypes; phenotype-token distributions are recovered *and* in the correct order.

which can leverage data subsampling and stochastic gradients may be useful. In addition, though our experiments with simulated data demonstrate SS3M is capable of recovering ground truth phenotypes, it is also true that the data generating distribution and the model belong to the same family. This match works to the benefit of the model, and we may expect the recovery of ground truth phenotypes to suffer when the data generating distribution belongs to a distinct family.

4.2 Aim 2B. *Utilize SS3M to learn disease-specific phenotypes from partially labeled observational clinical data.*

Background

The previous subaim describes SS3M and demonstrates it's ability to recover ground truth phenotypes in simulation. We now interrogate SS3M's ability to learn phenotypes from partially labeled clinical data. We train SS3M on partially labled clinical data extracted from the MIMIC-III critical care database. We also train a closely related, fully unsupservised baseline on the same data, but without the labels. We recruit two clinical experts to evaluate the quality and content of phenotypes learned by both models. In addition, we evaluate SS3M's predictive performance relative to common supervised baselines, and explore how this performance is impacted by the amount of labeled data made available during training.

Research Questions

Do SS3M phenotypes capture the clinical characteristics of the diseases specified by the labels provided?

Is SS3M an effective model for determining which patients have a specific disease?

Methods

Data. We train all our models using clinical data extracted from the Medical Information Mart for Intensive Care version III (MIMIC-III) (Johnson et al. 2016). Our dataset is restricted to adult patients where adults are defined as patients who are 18 years of age or older upon admission. Age upon admission is calculated by subtracting each patient's recorded date of birth from their time of admission. This constraint yields a cohort of 38,549 individual patients.

For each patient in our cohort, we extract observations from clinical notes, labs and medications. We refer to these data types as "data sources." Notes were restricted to

the following types: “Physician”, “General”, and “Discharge Summary”; no restrictions were placed on clinical labs and medications.

Each patient is represented by multiple sets of clinical observations (one set per data source) and a set of labels (possibly empty). We limit ourselves to the clinical observations and labels associated with each patient’s first hopsital admission.

In this work, we lack a set of true, expert-generated, gold-standard disease labels for our patient cohort. For this reason we make use of readily availble ICD9 diagnosis codes to contstruct our label set. Our labels correspond to a variety of disease conditions from the single-level definitions of the Health Cost and Utilization (HCUP) Clinical Classification Software (CCS). The HCUP CCS conditions are defined by groups of related ICD9 diagnosis codes. Relative to raw ICD9 codes, HCUP CCS code groups are significantly less noisy, which makes them attractive for phenotype prediction tasks in the absence of a true gold-standard. We apply all HCUP CCS single-level definitions to the ICD9 codes for our cohort and consider conditions with a least 10^3 positive cases (prevalence $\approx 2.5\%$). As MIMIC-III is a critical care database, we further limit ourselves to well represented acute conditions. This process led us to retain a total of 40 conditions for use in our experiments (See Table 4.3 and Figure 4.5 for our full list of HCUP CCS conditions). For each patient, we record a binary label for each of these disease conditions specifying its presence (1) or absence (0). We treat this label set as ground truth.

For a given patient, we concatenate all associated clinical observations within each data source. These observations are tokenized to yield a patient’s raw token representation in terms of words (from notes), lab names and medication names.

Tokenized notes are further preprocessed to remove English stop words as well as any word token with 20 appearances or less over the entire notes corpus. This latter step is intended to a filter out the large quantity of misspelled words observed in the unfiltered token vocabulary.

The notes vocabularly is further constrainted by applying a term-frequency/inverse-

document-frequence (TF-IDF) filter. For each patient d, and each token t observed in their tokenized set of notes we calculate a tf-idf weight w_{dt} as follows.

$$w_{dt} = N_{dt} \log_2 \frac{D}{N_t}, \tag{4.18}$$

where N_{dt} is the number of times token t appears in patient d's tokenized notes, N_d is the total number of patient d's note tokens, and D is the total number of patients. Next, we average these weights over all patients and retain the top 10^4 mean weighted tokens. No additional preprocessing was applied to clinical labs and medications post-tokenization.

We are interested in evaluating SS3M's ability to learn clincally meaningful phenotypes and perform phentoype prediction on held-out patient data. Moreover, we aim to evaluate SS3M's performance in these tasks when trained on various proportions of labeled patient data.

In the present setting, each patient has a full set of binary labels for each of our 40 HCUP CCS condition targets. These labels are treated as ground truth, and we train SS3M with subsets of them. To obtain each subset, we first specify a percentage of the training cohort for which we wish to retain labels. We then sample the corresponding number of patients from the training cohort and ensure the prevalence of each label in the labeled subset is similar to that in the total training cohort. During training, we use the full set of labels for each patient in the labeled subset. We carry out this process for various percentages of the training cohort including 1%, 5%, 25%, 50%, 75%, and 100%.

As described in Aim 2A, SS3M handles both semi-supervised and unsupervised phenotypes. In preliminary experiments, we observed SS3M's performance depended in part on the total number of phenotypes modeled, P. To characterize this dependency, we train SS3M on the label subsets described above with P set to 40 (i.e. no unsupervised phenotypes), 80 or 160.

Our total training cohort is comprised of 60% of the patient cohort described above. The

remaining 40% is reserved for validation (20%) and testing (20%).

Qualitative evaluation. Here we ask clinical experts to asses the quality of SS3M phenotypes relative to phenotypes inferred with a Multi-Channel Mixed Membership Model (MC3M), a closely related unsupervised model developed for phenotype inference (Pivovarov et al. 2015b). Like SS3M, MC3M learns mulitple phenotypes jointly from multi-source clinical data. We implement a collapsed Gibbs sampler for MC3M, and run inference on note, lab and medication data for the full training cohort.

We evaluate the quality of phenotypes learned with each model along three axis: *coherence*, *granularity*, and *label quality*. These axis and the methods for their evaluation are detailed in Pivovarov et al.

- *Coherence.* A coherent phenotype is defined as a phenotype containing observations typical of a single disease while omitting observations atypical of said disease. The clinical expert was asked to rate the coherence of individual phenotypes using a five-point Likert scale, with 1 and 5 signifying low and and high coherence, respectively.

- *Granularity.* Phenotype granularity is defined in terms of three categories: *(1)* non-disease, *(2)* mixture of diseases, *(3)* single disease. We asked our expert to assign each phenotype to one of these categories.

- *Label quality.* We asked our clinical expert to generate a label for each phenotype. If no such label came to mind, the expert was asked to omit this step. If the phenotype in question was learned using SS3M, the expert was asked if their label was equivalent to the phenotype's true label. In addition, the expert was asked to specify how well the true label matched its learned phenotype using a five-point Likert scale with 1 indicating no match and 5 a perfect match.

The phenotypes for our qualitative evaluations are learned using SS3M and MC3M models with $P = 160$. For SS3M, we use phenotypes learned using a labeled subset containing 75%

of the training cohort. Individual phenotypes from each model are visualized as sets of three word clouds, one for each data source (See Figures 4.4 and 4.5). Word clouds are generated using the WordCloud Python library (Mueller 2019).

We collaborate with two clinical experts to carry out our evaluation. Both evaluators are medical doctors who have completed or are near completing residency training in internal medicine.

To set up our evaluation we first randomly mix together the individual visualizations of the 40 semi-supervised SS3M phenotypes and 40 randomly chosen MC3M phenotypes, making sure to anonymize their model of origin. These visualizations are then given separately to each clinical expert along with a set of instructions. Each evaluator is also provided a spreadsheet for recording their evaluations. This spreadsheet specifies the order in which phenotypes are to be evaluated, and, for SS3M phenotypes, contains all the ground truth phenotype labels. Where applicable, we ensure evaluators are not exposed to a phenotype's ground truth label until they have completed its granularity and coherence assessments and suggested their own expert label.

We aggregate evaluations from each of our clinical experts and use Cohen's Kappa to calculate their interrater reliability within each evaluative task.

Quantitative evaluation. For each labeled subset and value of P, we obtain posterior estimates of SS3M's global latent variables ($\boldsymbol{B}$, B^*, $\boldsymbol{C}$, and $\boldsymbol{\Phi}$) by running our collapsed Gibbs sampler on the training data. These global variables are then passed to untrained SS3M models for which we run a partially collapsed Gibbs sampler (only $\boldsymbol{\Theta}$ is integrated out of the joint distribution) over the local latent variables ($\boldsymbol{A}$, $\boldsymbol{W}$, and $\boldsymbol{Z}$) on the test set. Within the held out set, the complete conditional likelihoods on each activation ($\boldsymbol{A}_{dp}$) are used as label prediction probabilities which we evaluate using the areas under the receiver operating characteristic and precision-recall curves (AUC-ROC, AUC-PR).

We compare SS3M's predictive performance to that of several commonly used baselines.

These include k-nearest neighbors (KNN) and random forests (RF), which we train as multilabel classifiers. We also compare against L1-regularized logistic regression (LR) trained as a set of 40 one-versus-rest classifiers, one for each target. Unlike, SS3M, our baselines were not developed to handle partially labeled training data. Thus, for any given configuration of the training cohort, we train baselinee on data for only those patients whose labels are included within the labeled subset.

Performance curves and baselines are estimated using the Scikit-learn Python library (Pedregosa et al. 2011a).

Results & Discussion

Qualitative Results. Figure 4.3 summarizes the results of our qualitative evaluation. On average, SS3M outperforms MC3M in terms of phenotype coherence and granularity. Over 90% of SS3M semi-supervised phenotypes showed high coherence (scores of 4 or 5) and nearly 80% were considered to have single-disease granularity. Meanwhile, unsupervised MC3M phenotypes had a more uniform distribution over all levels of coherence and granularity. In terms of label quality, about 75% of SS3M phenotypes were found to match well with their ground truth labels (scores of 4 or 5). Noteably, for nearly 80% of SS3M phenotypes, our expert evaluators were able to suggest a label that matched the ground truth label. This finding suggests that the large majority of SS3M semi-supervised phenotypes communicated the characteristics of the conditions described by their ground truth labels. Over all evaluative tasks, our expert evaluators demonstrate a fair degree of interrater reliability. Figure 4.4 displays a sample of phenotypes which received strong qualitative evaluations from both expert reviewers. Figure 4.5 shows the full set of semi-supervised phenotypes employed in the qualitative evaluation.

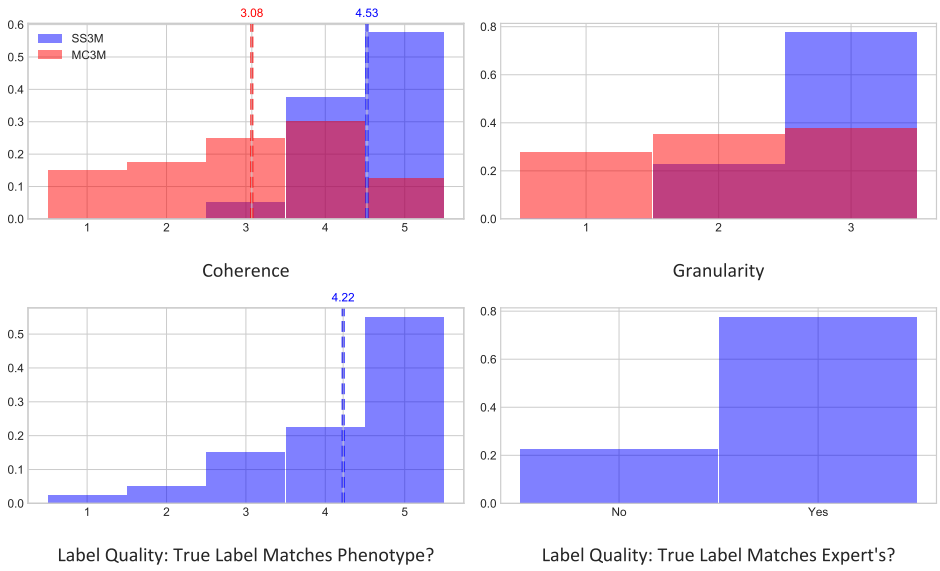

Figure 4.3: Qualitative evaluation results. Evaluator responses are aggregated within each evaluation type. Shown are the proportions of each possible response (as defined in Section 4.2). Means, where appropriate, are shown with vertical hashed lines. Interrater reliabilities (Cohen's κ): Coherence - 0.28; Granularity - 0.14; True label matches phenotype? - 0.04; True label matches expert's? - 0.50.

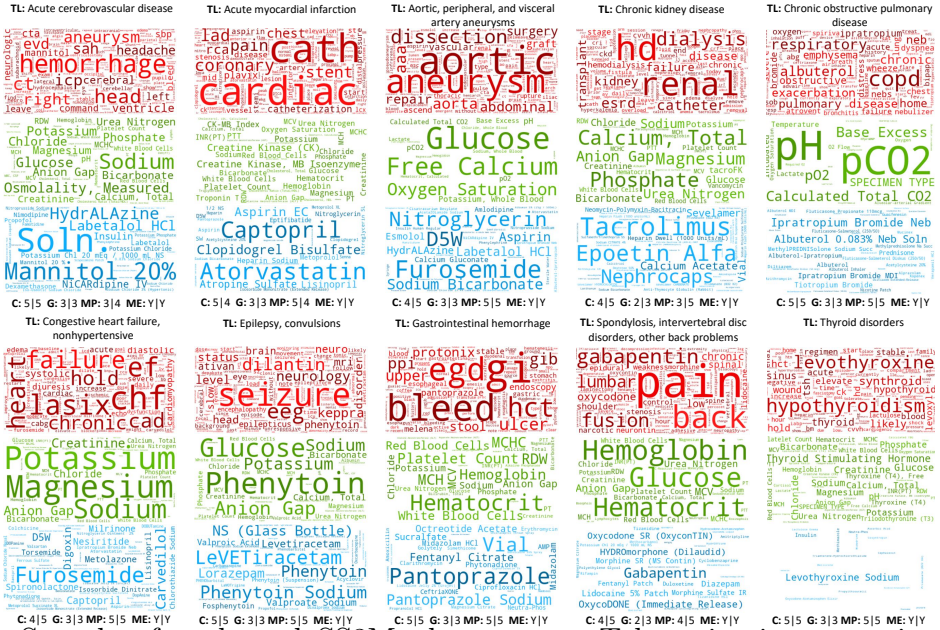

Figure 4.4: Sample of evaluated SS3M phenotypes. Token size is proportional to token likelihood within a phenotype. Red - words from clinical notes; Green - clinical lab names; Blue - medication names. Evaluations from both clincal experts are presented below each phenotype. **TL** - True label; **C** - Coherence; **G** - Granularity; **MP** - True label matches phenotype?; **ME** - True label matches expert's?

Figure 4.5: SS3M semi-supervised phenotypes. Phenotypes learned with P=160, and 75% labels retained for training. Token size is proportional to token likelihood within a phenotype. Red - words from clinical notes; Green - clinical lab names; Blue - medication names.

Quantitative Results. Table 4.2 summarizes the results of our quantitative evaluation. In general, SS3M's phenotype prediction performance, as measured by macro and micro averaged AUC-ROC and AUC-PR, grows for all values of P as the percentage of labeled patients increases from 1% to 100% of the training cohort. Moreover, performance appears to increase as P increases, particularly for larger amounts of labeled training data.

SS3M demonstrated competitive predictive performance relative to our baselines. In nearly all cases, SS3M with $P \geq 80$ outperforms our multilabel classification baselines (RF and KNN) once 25% of total labels are made available for training. In all cases, the set of 40 one-versus-rest L1-regularized logistic regression (LR) models outperformed all competitors. However, SS3M was the only multilabel classifier that approached LR's performance in at least a subset of cases (e.g. micro averaged AUC-ROC for $P = 160$ and 100% training labels).

Table 4.3 illustrates SS3M's per-label predictive performance for various proportions of labeled training data. As with the averaged predictive performance, per-label predictive performance tends to increase as more labels are made available for training. However, this trend is not entirely consistent. For some labels, performance increases for a time with the percentage of training labels, but then suddenly suffers a steep drop, possibly followed by a similarly steep rise. This volatility may be due in part to SS3M's MCMC inference algorithm which may get caught in similar but distinct posterior modes.

Average	Model	AUC-ROC (Training label %)						AUC-PR (Training label %)					
		1%	5%	25%	50%	75%	100%	1%	5%	25%	50%	75%	100%
Macro	SS3M (P=40)	0.557	0.653	0.723	0.73	0.723	0.737	0.156	0.22	0.29	0.305	0.286	0.302
	SS3M (P=80)	0.48	0.622	0.717	0.766	0.787	0.802	0.117	0.226	0.313	0.381	0.389	0.401
	SS3M (P=160)	0.445	0.54	0.702	0.781	0.798	0.813	0.0955	0.162	0.331	0.412	0.444	0.464
	RF (ML)	0.643	0.687	0.721	0.734	0.744	0.75	0.171	0.206	0.246	0.269	0.281	0.291
	KNN (ML)	0.605	0.641	0.679	0.695	0.701	0.704	0.15	0.184	0.228	0.246	0.256	0.263
	LR (OVR)	0.711	0.812	0.843	0.844	0.846	0.846	0.336	0.471	0.526	0.531	0.53	0.53
Micro	SS3M (P=40)	0.627	0.676	0.76	0.783	0.787	0.804	0.143	0.188	0.233	0.24	0.266	0.304
	SS3M (P=80)	0.629	0.699	0.786	0.837	0.847	0.858	0.18	0.241	0.329	0.431	0.442	0.465
	SS3M (P=160)	0.621	0.658	0.787	0.841	0.858	0.866	0.157	0.187	0.341	0.441	0.478	0.521
	RF (ML)	0.716	0.751	0.779	0.788	0.796	0.8	0.239	0.296	0.345	0.369	0.38	0.389
	KNN (ML)	0.657	0.693	0.725	0.74	0.743	0.746	0.193	0.239	0.288	0.309	0.319	0.325
	LR (OVR)	0.766	0.842	0.864	0.865	0.867	0.867	0.407	0.533	0.576	0.578	0.576	0.572

Table 4.2: Quantitative evaluation summary. Macro and micro averages are calculated for each model over all label targets. ML - multilabel classifier; OVR - one-versus-rest classifier.

	Prevalence		AUC-ROC (Training label %)						AUC-PR (Training label %)					
Label	Train (full)	Test	1%	5%	25%	50%	75%	100%	1%	5%	25%	50%	75%	100%
Acute and unspecified renal failure	0.205	0.207	0.505	0.483	0.808	0.569	0.841	0.856	0.242	0.246	0.64	0.362	0.681	0.671
Acute cerebrovascular disease	0.0841	0.0859	0.474	0.567	0.928	0.941	0.943	0.939	0.0834	0.123	0.599	0.711	0.714	0.744
Acute myocardial infarction	0.117	0.119	0.555	0.53	0.859	0.846	0.87	0.907	0.141	0.127	0.642	0.587	0.656	0.701
Acute posthemorrhagic anemia	0.0869	0.0939	0.616	0.484	0.499	0.729	0.746	0.766	0.141	0.108	0.107	0.311	0.367	0.379
Alcohol-related disorders	0.0866	0.0892	0.445	0.836	0.884	0.902	0.899	0.902	0.0811	0.563	0.636	0.652	0.68	0.665
Aortic, peripheral, and visceral artery aneurysms	0.0441	0.0419	0.502	0.434	0.888	0.907	0.906	0.871	0.0427	0.0344	0.52	0.459	0.545	0.5
Aspiration pneumonitis, food/vomitus	0.0706	0.0687	0.458	0.81	0.811	0.82	0.729	0.45	0.0639	0.268	0.39	0.271	0.219	0.0722
Asthma	0.0634	0.0659	0.478	0.475	0.906	0.891	0.894	0.898	0.0624	0.0634	0.383	0.349	0.371	0.354
Bacterial infection, unspecified site	0.0863	0.09	0.346	0.446	0.406	0.689	0.631	0.505	0.0684	0.0998	0.0822	0.254	0.249	0.142
Cardiac arrest and ventricular fibrillation	0.0339	0.0362	0.456	0.545	0.666	0.916	0.908	0.909	0.0389	0.055	0.104	0.396	0.323	0.298
Cardiac dysrhythmias	0.321	0.32	0.605	0.826	0.541	0.802	0.853	0.86	0.486	0.785	0.401	0.763	0.821	0.825
Chronic kidney disease	0.104	0.103	0.389	0.649	0.652	0.721	0.747	0.663	0.084	0.302	0.303	0.36	0.391	0.329
Chronic obstructive pulmonary disease and bronchiectasis	0.117	0.116	0.5	0.445	0.861	0.857	0.858	0.816	0.122	0.104	0.511	0.509	0.523	0.444
Coagulation and hemorrhagic disorders	0.109	0.102	0.332	0.671	0.447	0.732	0.756	0.757	0.0739	0.271	0.106	0.368	0.39	0.384
Congestive heart failure, nonhypertensive	0.241	0.233	0.545	0.45	0.832	0.79	0.805	0.85	0.328	0.214	0.727	0.655	0.688	0.745
Crushing injury or internal injury	0.0409	0.0424	0.513	0.604	0.903	0.873	0.877	0.895	0.0475	0.0633	0.456	0.457	0.481	0.538
Delirium, dementia, and amnestic and other cognitive disorders	0.0703	0.0687	0.394	0.419	0.83	0.855	0.871	0.874	0.0538	0.0575	0.448	0.434	0.369	0.419
Diabetes mellitus with complications	0.0757	0.08	0.444	0.401	0.932	0.942	0.935	0.918	0.068	0.0614	0.567	0.597	0.599	0.619
Epilepsy, convulsions	0.0643	0.063	0.448	0.917	0.933	0.92	0.931	0.919	0.0557	0.556	0.575	0.582	0.645	0.59
Gastrointestinal hemorrhage	0.0686	0.0732	0.407	0.472	0.921	0.914	0.902	0.897	0.0583	0.0775	0.579	0.559	0.612	0.611
Heart valve disorders	0.151	0.153	0.528	0.614	0.543	0.678	0.833	0.834	0.159	0.237	0.195	0.33	0.618	0.668
Hepatitis	0.0467	0.0482	0.417	0.746	0.858	0.794	0.842	0.801	0.0418	0.289	0.316	0.297	0.338	0.335
Intracranial injury	0.0633	0.0601	0.572	0.766	0.784	0.931	0.938	0.936	0.0811	0.185	0.257	0.552	0.595	0.565
Mood disorders	0.101	0.106	0.424	0.433	0.694	0.788	0.432	0.857	0.0896	0.0955	0.298	0.418	0.0972	0.486
Mycoses	0.0327	0.0358	0.301	0.425	0.552	0.481	0.572	0.559	0.0254	0.0378	0.073	0.0533	0.0906	0.0905
Open wounds of head, neck, and trunk	0.0283	0.0246	0.522	0.605	0.569	0.912	0.919	0.916	0.0294	0.0381	0.0323	0.236	0.252	0.235
Pancreatic disorders (not diabetes)	0.0283	0.0283	0.308	0.464	0.69	0.945	0.961	0.952	0.0192	0.032	0.216	0.416	0.502	0.426
Paralysis	0.0248	0.0298	0.39	0.44	0.533	0.555	0.655	0.585	0.0232	0.0267	0.0423	0.06	0.117	0.0756
Phlebitis, thrombophlebitis and thromboembolism	0.0584	0.0585	0.391	0.385	0.45	0.468	0.478	0.804	0.0494	0.0508	0.0581	0.0676	0.0727	0.35
Pleurisy, pneumothorax, pulmonary collapse	0.0949	0.101	0.432	0.495	0.622	0.69	0.679	0.616	0.0966	0.12	0.285	0.371	0.35	0.278
Pneumonia	0.133	0.142	0.348	0.48	0.754	0.778	0.791	0.81	0.111	0.171	0.498	0.509	0.506	0.541
Pulmonary heart disease	0.0635	0.0629	0.415	0.439	0.669	0.618	0.691	0.676	0.0519	0.0569	0.186	0.221	0.294	0.262
Respiratory failure, insufficiency, arrest (adult)	0.219	0.211	0.369	0.466	0.733	0.773	0.587	0.792	0.179	0.246	0.496	0.547	0.421	0.642
Secondary malignancies	0.0611	0.0611	0.463	0.482	0.958	0.959	0.959	0.958	0.0617	0.0616	0.617	0.647	0.625	0.615
Septicemia (except in labor)	0.136	0.133	0.362	0.549	0.534	0.705	0.506	0.778	0.111	0.216	0.204	0.422	0.183	0.52
Shock	0.0801	0.0791	0.406	0.583	0.531	0.852	0.876	0.859	0.0783	0.134	0.15	0.447	0.498	0.458
Spondylosis, intervertebral disc disorders, other back problems	0.0448	0.0468	0.439	0.478	0.721	0.748	0.761	0.804	0.0432	0.0505	0.211	0.226	0.228	0.254
Substance-related disorders	0.0419	0.0409	0.455	0.441	0.466	0.609	0.848	0.792	0.0375	0.0376	0.0383	0.101	0.288	0.3
Thyroid disorders	0.103	0.105	0.431	0.497	0.47	0.952	0.947	0.953	0.0895	0.11	0.105	0.664	0.691	0.68
Urinary tract infections	0.125	0.123	0.472	0.436	0.483	0.435	0.833	0.843	0.129	0.114	0.134	0.116	0.558	0.544
Macro Average			0.445	0.54	0.702	0.781	0.798	0.813	0.0955	0.162	0.331	0.412	0.444	0.464
Micro Average			0.622	0.658	0.788	0.842	0.858	0.866	0.158	0.188	0.341	0.44	0.476	0.518

Table 4.3: SS3M per-label predictive performance. Shown are results for $P = 160$, and 75% labels retained for training.

Limitations

SS3M's predictive performance is competitive, but not clearly superior to that of baselines. This is not surprising; discriminitive models are known to outperform generative models when training data are large (Ng and Jordan 2002); however, a more performant model would be desirable. Furthermore, our experiments are limited by our choice of labels. Since gold-standard labels are not available, we use aggregated HCUP CCS codes to identify cases for each of our target diseases. These aggregated codes are less noisy than raw diagnosis codes, but they are, at best, a silver standard.

Chapter 5

Aim 3. Develop a method for learning phenotypes from partially observed, partially labeled data.

Interest in modeling clinical data stores has surged in recent years as electronic capture of clinical observations continues to expand. Nevertheless, clinical data, by its nature, presents significant modeling challenges which most machine learning algorithms are not suited to overcome. These challenges derive from the significant data quality issues which persistently characterize observational health data (Weiskopf and Weng 2013; Weiskopf et al. 2013). Clinical observations are made and recorded in a manner consistent with efficacious health care delivery; they are generally not gathered meticulously, or systematically for secondary use (Hripcsak, Albers, and Perotte 2015). This observation alone points to the chasm that exists between the research grade data-sets commonly used within the general machine learning (ML) community and the data that emerges from the healthcare system. This is before we consider the sporadic nature of clinical encounters, and the limited tools available for articulating patient state when such encounters do occur. In fact, even under the most vigilant clinical scenarios, we would be unlikely to capture the full spectrum of clinically informative observations for a patient, let alone record them for large patient cohorts on a longitudinal basis. (Newton et al. 2013)

Since clinical observations often occur infrequently, noisily, or not at all, missingness is salient in most clinical datasets. Furthermore, as outlined above, this characteristic is likely to persist in healthcare data for the foreseeable future. Therefore, it is crucial for the

machine learning in healthcare community to develop models and inference algorithms that can handle missing data with robustness.

Among the most exciting and broadly applicable research areas in modern machine learning is the work on deep generative models such as Variational Autoencoders (VAE) (Rezende, Mohamed, and Wierstra 2014; Kingma and Welling 2014). Much of this work has focused on modeling complete, dense, richly correlated image data-sets. However, in recent years several authors have begun adapting these frameworks specifically to handle lower-dimensional data with missing values. Among these approaches, some rely upon access to fully observed training data (Rezende, Mohamed, and Wierstra 2014), while others fill missing values with suitable placeholders during training (Nazábal et al. 2020; Mattei and Frellsen 2019; Ipsen, Mattei, and Frellsen 2021). Ideally, a deep generative model for use on healthcare data would assume partially observed training data. In addition, a model built to focus specifically on the pure observed data signal would be beneficial, as any artificial signal injected into the training routine may distract our models from the learning task at hand.

In this work we introduce a simple Monte Carlo Expectation-Maximization (MCEM) algorithm for training VAEs on partially observed data. Our model does not require complete data to train, nor does it rely on placeholder values for missing data. To illustrate the utility of our approach, we evaluate VAEs trained with our MCEM algorithm on a set of data imputation tasks. We compare against several baseline imputers, and demonstrate our method consistently produces superior results. We then apply our method in a semi-supervised setting where the data contain partially observed binary labels. We evaluate how well the resultant models predict missing labels using standard supervised learning metrics. In experiments with clinical data, labels correspond to binary patient states (e.g. patient is intubated, patient has type II diabetes). In this case, our approach yields a probabilistic phenotyping algorithm based on multi-label classification.

5.1 Aim 3A. *Derive and implement a Monte Carlo Expectation-Maximization (MCEM) algorithm for training Variational Autoencoders (VAEs) on partially observed data.*

Background

Clinical data often contain missing values, which presents a challenge when using deep generative modeling frameworks like Variational Autoencoders (VAEs) that assume fully observed data during training. To address this issue, we develop an algorithm for training VAEs on partially observed clinical data. Our approach employs a Monte Carlo Expectation-Maximization (MCEM) algorithm which maximizes a double lower bound on the marginal log-likelihood of the observed data.

We evaluate the performance of VAEs trained using our algorithm on a set of missing value imputation tasks. In our experiments, we compare our method to other deep generative imputers, which we consistently outperform.

Research Question

Is a VAE trained with MCEM an effective model for performing missing value imputation?

Methods

Model. In this research we consider VAEs as described by Kingma and Welling 2014. Assume a dataset $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^N$ consisting of N i.i.d. samples from some unspecified data generating distribution. The VAE models the data along with a set of latent variables $\boldsymbol{Z} = \{\boldsymbol{z}_i\}_{i=1}^N$ using a joint distribution $p_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Z}) = \prod_{i=1}^N p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{z}_i)p(\boldsymbol{z}_i)$ parameterized by a neural network $\boldsymbol{\theta}$. Because posterior inference in this model is intractable, VAEs introduce a variational approximation to the posterior $q_{\boldsymbol{\phi}}(\boldsymbol{Z}|\boldsymbol{X}) = \prod_{i=1}^N q_{\boldsymbol{\phi}}(\boldsymbol{z}_i|\boldsymbol{x}_i)$ where $\boldsymbol{\phi}$ are neural

network parameters. The parameters $\{\boldsymbol{\theta}, \boldsymbol{\phi}\}$ are optimized jointly using Auto-Encoding Variational Bayes (AEVB), an inference algorithm we discuss further below. Hereafter we drop instance indices wherever context renders them unnecessary.

Inference. As originally proposed, VAEs assume fully observed data; however, often the data are only partially observed. In this case, each instance may be partitioned into observed and missing components: $\boldsymbol{x} = (\boldsymbol{x}_o, \boldsymbol{x}_m)$ where $\boldsymbol{x}_o = \{x \in \boldsymbol{x} : x \text{ is } \textit{observed}\}$ and $\boldsymbol{x}_m = \{x \in \boldsymbol{x} : x \text{ is } \textit{missing}\}$. Our goal is to develop a simple inference algorithm to train VAEs in this setting. To do so we will show how AEVB may be incorporated into an MCEM algorithm which maximizes a lower bound on the marginal log-likelihood of the observed data, $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_o)$. We begin by reviewing Expectation-Maximization (EM).

Expectation-Maximization. EM is an iterative algorithm used to find maximum likelihood solutions for models containing observed variables, $\boldsymbol{x}$, unobserved (i.e. latent or missing) variables $\boldsymbol{z}$, and parameters, $\boldsymbol{\theta}$. Each iteration includes an expectation step (E-step) and a maximization step (M-step) (Dempster, Laird, and Rubin 1977).

$$\text{E-step:} \quad Q_{\text{EM}}(\boldsymbol{\theta}', \boldsymbol{\theta}) = \mathbb{E}_{p_{\boldsymbol{\theta}'}(\boldsymbol{z}|\boldsymbol{x})} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})\right] \tag{5.1}$$

$$\text{M-step:} \quad \boldsymbol{\theta}'' = \operatorname*{argmax}_{\boldsymbol{\theta}} Q_{\text{EM}}(\boldsymbol{\theta}', \boldsymbol{\theta}) \tag{5.2}$$

Here, $\boldsymbol{\theta}'$ is the value of the $\boldsymbol{\theta}$ at the end of the previous iteration, while $\boldsymbol{\theta}''$ is the value of $\boldsymbol{\theta}$ at the end of the current iteration.

We can derive the EM algorithm following Bishop 2006. Our goal is to maximize the marginal log-likelihood, $\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$, which is also called the log evidence. To do so, we first calculate an expectation w.r.t. an auxiliary distribution over the latent variables, $q(\boldsymbol{z})$.

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{z})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})}{p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})}\right] = \mathbb{E}_{q(\boldsymbol{z})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})}\right] + \mathbb{E}_{q(\boldsymbol{z})}\left[\log \frac{q(\boldsymbol{z})}{p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})}\right] \tag{5.3}$$

$$= \mathcal{L}_{\text{EM}}(\boldsymbol{\theta}; \boldsymbol{x}) + \text{KL}(q(z) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})) \tag{5.4}$$

Because $\mathrm{KL}(q(z) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})) \geq 0$, it must be the case that the first term, $\mathcal{L}_{\mathrm{EM}}(\boldsymbol{\theta}; \boldsymbol{x})$, lower bounds the log evidence. Hence, it is often referred to as the Evidence Lower BOund or the ELBO.

Note that setting $q(\boldsymbol{z})$ equal to the posterior $p_{\boldsymbol{\theta}'}(\boldsymbol{z}|\boldsymbol{x})$ yields an ELBO which is equivalent to the EM objective, Q_{EM} (up to an entropy term which has no $\boldsymbol{\theta}$-dependence).

$$\mathcal{L}_{\mathrm{EM}}(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{\theta}') = \mathbb{E}_{p_{\boldsymbol{\theta}'}(\boldsymbol{z}|\boldsymbol{x})} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})\right] + \mathbb{H}_{\boldsymbol{\theta}'}[\boldsymbol{z}|\boldsymbol{x}] = Q_{\mathrm{EM}}(\boldsymbol{\theta}', \boldsymbol{\theta}) + \text{const.} \tag{5.5}$$

Thus, maximizing Q_{EM} w.r.t. $\boldsymbol{\theta}$ also maximizes the ELBO, $\mathcal{L}_{\mathrm{EM}}$. Maximizing $\mathcal{L}_{\mathrm{EM}}$, in turn, maximizes the marginal log-likelihood, $\log p_{\boldsymbol{\theta}}(x)$ which the ELBO lower bounds.

Monte Carlo Expectation-Maximization. MCEM uses a Monte Carlo estimate to approximate the expectation in the EM algorithm's E-step.

$$\text{E-step:} \quad \hat{Q}_{\mathrm{MCEM}}(\boldsymbol{\theta}', \boldsymbol{\theta}) = \sum_{l=1}^{L} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}^{(l)}) \; ; \;\; \boldsymbol{z}^{(l)} \sim p_{\boldsymbol{\theta}'}(\boldsymbol{z}|\boldsymbol{x}) \tag{5.6}$$

$$\text{M-step:} \quad \boldsymbol{\theta}'' = \operatorname*{argmax}_{\boldsymbol{\theta}} \hat{Q}_{\mathrm{MCEM}}(\boldsymbol{\theta}', \boldsymbol{\theta}) \tag{5.7}$$

MCEM is useful when calculating expectations w.r.t. the posterior is difficult but sampling from it is simple.

Variational Inference. VI is closely related to EM (Blei, Kucukelbir, and McAuliffe 2017; Bishop 2006). In VI, the auxiliary distribution is replaced with a variationl distribution, $q_{\boldsymbol{\phi}}(\boldsymbol{z})$, embued with its own parameters, $\boldsymbol{\phi}$. The ELBO in this case may be written as follows.

$$\mathcal{L}_{\mathrm{VI}}(\boldsymbol{\phi}; \boldsymbol{x}, \boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \mathrm{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})) \tag{5.8}$$

Since the marginal log-likelihood has no $\boldsymbol{\phi}$-dependence, maximizing the ELBO w.r.t. $\boldsymbol{\phi}$ necessarily minimizes the KL term. Thus, the variational parameters, $\boldsymbol{\phi}$, are optimized such

that $q_{\boldsymbol{\phi}}(\boldsymbol{z})$ approximates the posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$.

Auto-Encoding Variational Bayes. AEVB is a variant of VI in which the variational distribution is conditioned on the observed variables, $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$, and the likelihood and variational parameters, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, are assumed to be neural networks (Kingma and Welling 2014). Most commonly, the joint likelihood is assumed to factorize as follows $p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$. Here the "decoder", $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$, relies on the neural network paramters, $\boldsymbol{\theta}$, to ingest samples of the latent variable $\boldsymbol{z}$ and output distribution parameters appropriate for modeling $\boldsymbol{x}$ (e.g. location and scale for continuous $\boldsymbol{x}$ modeled using a normal distribution). Similarly, the "encoder", $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$, uses neural network parameters $\boldsymbol{\phi}$ to ingest $\boldsymbol{x}$ and generate distribution parameters for $\boldsymbol{z}$ (typically normally distributed). The prior, $p(\boldsymbol{z})$, is usually set to standard normal, though this framework permits priors with their own trainable parameters as we will discuss later. This configuration of distributions and parameters is commonly referred to as a Variational Autoencoder or VAE (Kingma and Welling 2014; Rezende, Mohamed, and Wierstra 2014). To fit the model, the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are optimized simultaneously to maximize the ELBO, $\mathcal{L}_{\text{VAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x})$:

$$\mathcal{L}_{\text{VAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] - \text{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z})) \tag{5.9}$$

Imputation with VAEs. Rezende, Mohamed, and Wierstra 2014 introduce a method for using a trained VAE to impute missing values. Assume we have trained VAE parameters $\{\boldsymbol{\theta}, \boldsymbol{\phi}\}$ using fully observed data. Now, we obtain a partially observed test instance, $\boldsymbol{x} = (\boldsymbol{x}_o, \boldsymbol{x}_m)$. To impute missing values, we randomly initialize $\boldsymbol{x}_m$, and push $\boldsymbol{x}$ through our trained VAE, overwriting $\boldsymbol{x}_m$ with samples from the VAE's likelihood. This process is iterated until the imputed values stabilize.

This procedure induces a Markov chain with the following transition kernel.

$$T_{\{\boldsymbol{\theta},\boldsymbol{\phi}\}}(\boldsymbol{x}'_m|\boldsymbol{x}_o, \boldsymbol{x}_m) = \int\int p_{\boldsymbol{\theta}}(\boldsymbol{x}'_o, \boldsymbol{x}'_m|\boldsymbol{z}) q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_o, \boldsymbol{x}_m)\, d\boldsymbol{x}'_o\, d\boldsymbol{z} \tag{5.10}$$

If $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ closely approximates $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$, then $T_{\{\boldsymbol{\theta},\boldsymbol{\phi}\}}$ closely approximates the following kernel

$$T_{\boldsymbol{\theta}}(\boldsymbol{x}'_m|\boldsymbol{x}_o,\boldsymbol{x}_m) = \int\int p_{\boldsymbol{\theta}}(\boldsymbol{x}'_o,\boldsymbol{x}'_m|\boldsymbol{z})p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_o,\boldsymbol{x}_m)\,d\boldsymbol{x}'_o\,d\boldsymbol{z}. \tag{5.11}$$

As Rezende, Mohamed, and Wierstra 2014 show, this kernel has $p_{\boldsymbol{\theta}}(\boldsymbol{x}_m|\boldsymbol{x}_o)$ as its equilibrium distribution. We will also show this to be true below, but, for now, it is worth pausing to consider this distribution. For the purposes of data imputation, this conditional is ideal; under our model, it represents everything we could possibly know about $\boldsymbol{x_m}$ given our knowledge of $\boldsymbol{x_o}$. As such, one could consider a "good" imputation model to be one which induces a $p_{\boldsymbol{\theta}}(\boldsymbol{x}_m|\boldsymbol{x}_o)$ that, in some way, approximates the true conditional $p(\boldsymbol{x}_m|\boldsymbol{x}_o)$ (e.g. by generating samples drawn, approximately, from $p(\boldsymbol{x}_m|\boldsymbol{x}_o)$). We will return this idea later when discussing our evaluation strategy for the current subaim.

As promised, we now show that $T_{\boldsymbol{\theta}}(\boldsymbol{x}'_m|\boldsymbol{x}_o,\boldsymbol{x}_m)$ has $p_{\boldsymbol{\theta}}(\boldsymbol{x}_m|\boldsymbol{x}_o)$ as its equilibrium distribution. To do so, we demonstrate that, when applied to $p_{\boldsymbol{\theta}}(\boldsymbol{x}_m|\boldsymbol{x}_o)$, $T_{\boldsymbol{\theta}}$ obeys the sufficient (but not necessary) detailed balance condition,

$$T_{\boldsymbol{\theta}}(\boldsymbol{x}'_m|\boldsymbol{x}_o,\boldsymbol{x}_m)p_{\boldsymbol{\theta}}(\boldsymbol{x}_m|\boldsymbol{x}_o) = T_{\boldsymbol{\theta}}(\boldsymbol{x}_m|\boldsymbol{x}_o,\boldsymbol{x}'_m)p_{\boldsymbol{\theta}}(\boldsymbol{x}'_m|\boldsymbol{x}_o), \tag{5.12}$$

or equivalently,

$$\int T_{\boldsymbol{\theta}}(\boldsymbol{x}'_m|\boldsymbol{x}_o,\boldsymbol{x}_m)p_{\boldsymbol{\theta}}(\boldsymbol{x}_m|\boldsymbol{x}_o)\,d\boldsymbol{x}_m = p_{\boldsymbol{\theta}}(\boldsymbol{x}'_m|\boldsymbol{x}_o). \tag{5.13}$$

We need only expand and evaluate the integral above:

$$\int T_{\boldsymbol{\theta}}(\boldsymbol{x}'_m|\boldsymbol{x}_o,\boldsymbol{x}_m)p_{\boldsymbol{\theta}}(\boldsymbol{x}_m|\boldsymbol{x}_o)\,d\boldsymbol{x}_m = \int\int\int p_{\boldsymbol{\theta}}(\boldsymbol{x}'_o,\boldsymbol{x}'_m|\boldsymbol{z})p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}_o,\boldsymbol{x}_m)p_{\boldsymbol{\theta}}(\boldsymbol{x}_m|\boldsymbol{x}_o)\,d\boldsymbol{x}'_o\,d\boldsymbol{x}_m\,d\boldsymbol{z} \tag{5.14}$$

$$= \int\int\int p_{\boldsymbol{\theta}}(\boldsymbol{x}'_o,\boldsymbol{x}'_m,\boldsymbol{x}_m,\boldsymbol{z}|\boldsymbol{x}_o)\,d\boldsymbol{x}'_o\,d\boldsymbol{x}_m\,d\boldsymbol{z} \tag{5.15}$$

$$= p_{\boldsymbol{\theta}}(\boldsymbol{x}'_m|\boldsymbol{x}_o). \tag{5.16}$$

Moreover, the smoothness of the VAE's likelihood guarantee $T_{\boldsymbol{\theta}}(\boldsymbol{x}'_m|\boldsymbol{x}_o, \boldsymbol{x}_m) > 0$ for any $\boldsymbol{x}'_m, \boldsymbol{x}_o, \boldsymbol{x}_m$ (Rezende, Mohamed, and Wierstra 2014). By the fundamental theorem for Markov chains, these two properties combine to ensure repeated sampling from $T_{\boldsymbol{\theta}}$ will eventually yield samples from $p_{\boldsymbol{\theta}}(\boldsymbol{x}_m|\boldsymbol{x}_o)$ (Neal 1993). Thus, repeated sampling from $T_{\{\boldsymbol{\theta},\boldsymbol{\phi}\}}$ will yield approximate samples from $p_{\boldsymbol{\theta}}(\boldsymbol{x}_m|\boldsymbol{x}_o)$ when $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \approx p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$.

Training VAEs on Partially Observed Data with MCEM. We now have all the pieces needed to describe our MCEM algorithm for training VAEs on partially observed data. Our objective is to maximize the log-likelihood of the observed data, $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_o)$; we employ the EM algorithm using an auxiliary distribution over the missing data, $q(\boldsymbol{x}_m)$.

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_o) = \mathbb{E}_{q(\boldsymbol{x}_m)}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_o, \boldsymbol{x}_m)}{p_{\boldsymbol{\theta}}(\boldsymbol{x}_m|\boldsymbol{x}_o)}\right] = \mathbb{E}_{q(\boldsymbol{x}_m)}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_o, \boldsymbol{x}_m)}{q(\boldsymbol{x}_m)}\right] + \mathbb{E}_{q(\boldsymbol{x}_m)}\left[\log \frac{q(\boldsymbol{x}_m)}{p_{\boldsymbol{\theta}}(\boldsymbol{x}_m|\boldsymbol{x}_o)}\right] \tag{5.17}$$

$$= \mathcal{L}_{\text{EM}}(\boldsymbol{\theta}; \boldsymbol{x}_o) + \text{KL}(q(\boldsymbol{x}_m) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{x}_m|\boldsymbol{x}_o)) \tag{5.18}$$

For the time being, let's assume we have access to the conditional distribution $p_{\boldsymbol{\theta}'}(\boldsymbol{x}_m|\boldsymbol{x}_o)$ and can calculate expectations w.r.t. it. We can set our auxillary distribution equal to this conditional to write $\mathcal{L}_{\text{EM}}$ as a function of the EM objective, Q_{EM}.

$$\mathcal{L}_{\text{EM}}(\boldsymbol{\theta}; \boldsymbol{x}_o, \boldsymbol{\theta}') = \mathbb{E}_{p_{\boldsymbol{\theta}'}(\boldsymbol{x}_m|\boldsymbol{x}_o)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_o, \boldsymbol{x}_m)\right] + \mathbb{H}_{\boldsymbol{\theta}'}[\boldsymbol{x}_m|\boldsymbol{x}_o] = Q_{\text{EM}}(\boldsymbol{\theta}', \boldsymbol{\theta}) + \text{const.} \tag{5.19}$$

Now, let's consider Q_{EM} in integral form.

$$Q_{\text{EM}}(\boldsymbol{\theta}', \boldsymbol{\theta}) = \int p_{\boldsymbol{\theta}'}(\boldsymbol{x}_m|\boldsymbol{x}_o) \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_o, \boldsymbol{x}_m)\, d\boldsymbol{x}_m\,. \tag{5.20}$$

We don't actually know $p_{\boldsymbol{\theta}'}(\boldsymbol{x}_m|\boldsymbol{x}_o)$, so estimating this integral is difficult. However, we can obtain (approximate) samples from $p_{\boldsymbol{\theta}'}(\boldsymbol{x}_m|\boldsymbol{x}_o)$ by using a VAE to construct the transition kernel, $T_{\{\boldsymbol{\theta},\boldsymbol{\phi}\}}$. The ELBO for this VAE may be obtained as a lower bound to the log-

likelihood of $\boldsymbol{x} = (\boldsymbol{x}_o, \boldsymbol{x}_m)$.

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_o, \boldsymbol{x}_m) = \log \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] \geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] \equiv \mathcal{L}_{\text{VAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_o, \boldsymbol{x}_m) \,, \tag{5.21}$$

where we've used Jensen's inequality to establish the bound. We can substitute this inequality into 5.20 to obtain a lower bound on Q_{EM}.

$$Q_{\text{EM}}(\boldsymbol{\theta}', \boldsymbol{\theta}) \geq \int p_{\boldsymbol{\theta}'}(\boldsymbol{x}_m|\boldsymbol{x}_o)\mathcal{L}_{\text{VAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_o, \boldsymbol{x}_m)\, d\boldsymbol{x}_m \equiv Q_{\text{EM-VAE}}(\boldsymbol{\theta}', \{\boldsymbol{\theta}, \boldsymbol{\phi}\}) \,, \tag{5.22}$$

Finally, we may obtain a Monte Carlo estimate of this integral with samples from $T_{\{\boldsymbol{\theta}', \boldsymbol{\phi}'\}}$.

$$Q_{\text{EM-VAE}}(\boldsymbol{\theta}', \{\boldsymbol{\theta}, \boldsymbol{\phi}\}) \approx \sum_{l=1}^{L} \mathcal{L}_{\text{VAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_o, \boldsymbol{x}_m^{(l)}) \,, \;\; \boldsymbol{x}_m^{(l)} \sim T_{\{\boldsymbol{\theta}', \boldsymbol{\phi}'\}}(\boldsymbol{x}'_m|\boldsymbol{x}_o, \boldsymbol{x}_m) \tag{5.23}$$

$$\equiv \hat{Q}_{\text{MCEM-VAE}}(\{\boldsymbol{\theta}', \boldsymbol{\phi}'\}, \{\boldsymbol{\theta}, \boldsymbol{\phi}\}) \tag{5.24}$$

This objective provides an accessible *double* lower bound to the observed marginal log-likelihood.

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_o) \geq Q_{\text{EM}}(\boldsymbol{\theta}', \boldsymbol{\theta} + \text{const.} \geq Q_{\text{EM-VAE}}(\boldsymbol{\theta}', \{\boldsymbol{\theta}, \boldsymbol{\phi}\}) + \text{const.} \tag{5.25}$$

$$\approx \hat{Q}_{\text{MCEM-VAE}}(\{\boldsymbol{\theta}', \boldsymbol{\phi}'\}, \{\boldsymbol{\theta}, \boldsymbol{\phi}\}) + \text{const.} \tag{5.26}$$

Thus, maximizing $\hat{Q}_{\text{MCEM-VAE}}$ within an MCEM algorithm maximizes $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_o)$ as was originally desired.

$$\text{E-step:} \quad \hat{Q}_{\text{MCEM-VAE}}(\{\boldsymbol{\theta}', \boldsymbol{\phi}'\}, \{\boldsymbol{\theta}, \boldsymbol{\phi}\}) = \sum_{l=1}^{L} \mathcal{L}_{\text{VAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_o, \boldsymbol{x}_m^{(l)}) \,, \;\; \boldsymbol{x}_m^{(l)} \sim T_{\{\boldsymbol{\theta}', \boldsymbol{\phi}'\}}(\boldsymbol{x}'_m|\boldsymbol{x}_o, \boldsymbol{x}_m) \tag{5.27}$$

$$\text{M-step:} \quad \{\boldsymbol{\theta''}, \boldsymbol{\phi''}\} = \underset{\{\boldsymbol{\theta}, \boldsymbol{\phi}\}}{\text{argmax}}\, \hat{Q}_{\text{MCEM-VAE}}(\{\boldsymbol{\theta}', \boldsymbol{\phi}'\}, \{\boldsymbol{\theta}, \boldsymbol{\phi}\}) \tag{5.28}$$

Note that maximizing the VAE paramters, $\{\boldsymbol{\theta}, \boldsymbol{\phi}\}$, in the M-step is straightforward; we just need to run AEVB which maximizes $\mathcal{L}_{\text{VAE}}$ and, therefore, $\hat{Q}_{\text{MCEM-VAE}}$ as well.

Sampling Importance Resampling. When the data are multimodal, multiple regions of the latent space may be needed to fully explain a given data point. This may be a common occurrence when the latent space may also be mutlimodal as we discuss later. However, Markov chains are known to have difficulty mixing across modes (Robert, Casella, and Casella 2010). This presents a challenge to our MCEM algorithm, which we address by incorporating sampling importance resampling (SIR) (Gelman et al. 1995; Bishop 2006). Briefly, for each data point, $\boldsymbol{x} = (\boldsymbol{x_o}, \boldsymbol{x_m})$, we sample from the prior, $p(\boldsymbol{z})$, to obtain L latent variables, $\{\boldsymbol{z}^{(l)}\}_{l=1}^{L}$, distributed throughout the latent space. We then use the conditional likelihood, $p(\boldsymbol{x}|\boldsymbol{z})$, to test how well each sample explains the observed data, $\boldsymbol{x}_o$. These likelihoods are then used to instantiate a resampling distribution over the latent samples. Samples that better explain the data are more likely to be resampled, regardless of which region of the latent space they originate from.

To use SIR, we first sample $L' \gg L$ latent variables and define a set of importance weights, $\{w_l\}_{l=1}^{L'}$, which take the form of self-normalized likelihood ratios between target and proposal distributions. We set the target and proposal to the joint, $p_{\boldsymbol{\theta}}(\boldsymbol{x}_o, \boldsymbol{z})$, and the prior, $p(\boldsymbol{z})$, respectively.

$$w_l = \frac{r_l}{\sum_{l=1}^{L'} r_l}, \quad r_l = \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_o, \boldsymbol{z^{(l)}})}{p(\boldsymbol{z^{(l)}})} = p_{\boldsymbol{\theta}}(\boldsymbol{x}_o|\boldsymbol{z^{(l)}}) \tag{5.29}$$

These importance weights are then used to instantiate a discrete resampling distribution over our latent variables, which we sample from with replacement to obtain a final set of L samples.

At each training interation, we use SIR to intialize missing values prior to running MCEM. First, we use SIR to obtain a set of L latent variables for each data point. Then, for each latent variable, $\boldsymbol{z}^{(l)}$, we sample $\boldsymbol{x}^{(l)} = (\boldsymbol{x}_o^{(l)}, \boldsymbol{x}_m^{(l)})$ from $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z} = \boldsymbol{z}^{(l)})$, and use the $\boldsymbol{x}_m^{(l)}$ to

instantiate the missing values. Finally, the full set of initialized samples, $\{\boldsymbol{x}_o, \boldsymbol{x}_m^{(l)}\}_{l=1}^L$, is run through MCEM.

VAE Architecture. We use multilayer perceptrons (MLPs) with ReLU nonlinearities in all VAE encoders and decoders. In addition, we find that adding trainable parameters to the prior on $\boldsymbol{z}$ often improves the quality of imputations. Thus, instead of modeling the prior as a fixed standard normal distribution, we use a normalizing flow (NF) (Papamakarios et al. 2021; Kobyzev, Prince, and Brubaker 2020). Briefly, a NF is a generative model which uses an invertible function, $T_{\boldsymbol{\xi}}$, with trainable parameters, $\boldsymbol{\xi}$, to transform a simple base distribution, $p(\boldsymbol{u}')$ (e.g. standard normal), into a more complex one, $p_{\boldsymbol{\xi}}(\boldsymbol{u})$. This transformation relies upon the change of variables formula:

$$p_{\boldsymbol{\xi}}(\boldsymbol{u}) = p(T_{\boldsymbol{\xi}}(\boldsymbol{u})) \left| \det \frac{\partial T_{\boldsymbol{\xi}}(\boldsymbol{u})}{\partial \boldsymbol{u}} \right| = p(\boldsymbol{u}') \left| \det \frac{\partial T_{\boldsymbol{\xi}}^{-1}(\boldsymbol{u}')}{\partial \boldsymbol{u}'} \right|^{-1}. \tag{5.30}$$

Usually, NFs are trained using maximum likelihood. In our case, the NF parameters are optimized alongside the encoder and decoder parameters using AEVB. The ELBO incorporates the flow parameters as follows:

$$\mathcal{L}_{\text{VAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\xi}; \boldsymbol{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p_{\boldsymbol{\theta},\boldsymbol{\xi}}(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) p_{\boldsymbol{\xi}}(\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \right]. \tag{5.31}$$

We note that previous works have explored modeling the prior, $p(\boldsymbol{z})$, using NFs (Xu et al. 2019). To date, these works have focused primarily on improving generative modeling with VAEs, and have not explored the utility of this kind of VAE for modeling partially observed data or missing value imputation.

The prior distribution influences the structure of the latent space via the negated KL term in $\mathcal{L}_{\text{VAE}}$: $\text{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}))$. When the prior is set to standard normal, this term functions as a regularizer encouraging $q_{\boldsymbol{\phi}}$ to push probability mass to high likelihood regions under standard normal. When the prior is modeled with an NF, this regularization effect

evolves during training as both the NF and encoder parameters are updated. This permits learning a latent space which can be more complex than standard normal and which may contain multiple modes. As such, the SIR algorithm we discuss above plays an important role when using an NF prior, as it allows our MCEM training scheme to leverage multiple latent modes to explain the observed data.

Generally, different NFs are defined based on their specification of $T_{\boldsymbol{\xi}}$. In the present work we experimented with Masked Autoregressive Flows (MAF) (Papamakarios, Pavlakou, and Murray 2018) and Real-Valued Non-Volume Preserving (RealNVP) (Dinh, Sohl-Dickstein, and Bengio 2022) flows. We find MAF consistently outperformed RealNVP, thus we use MAF throughout. In all cases, we use the standard normal as our base distribution.

Baselines. We compare our imputation method to two VAE-based imputers. The first of these is MVAE (Nazábal et al. 2020), which trains on partially observed data by simply initializing missing values to zero and maximizing a modified version of the ELBO in which the conditonal likelihood term is limited to only the observed subset of variables:

$$\mathcal{L}_{\text{MVAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_o) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_o, \boldsymbol{x}_m^0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_o|\boldsymbol{z}) p(\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_o, \boldsymbol{x}_m^0)} \right], \tag{5.32}$$

where we use $\boldsymbol{x}_m^0$ to denote zero-imputed missing values.

The second VAE-based baseine is the Missingness Importance Weighted AutoEncoder (MIWAE) (Mattei and Frellsen 2019). MIWAE generates multiple latent samples per data point to optimizes an importance weighted objective targeting the marginal likelihood of the observed data. Like MVAE, MIWAE uses zero-imputation to initialize missing values:

$$\mathcal{L}_{\text{MIWAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_o) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_o, \boldsymbol{x}_m^0)} \left[\log \frac{1}{L} \sum_{l=1}^{L} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_o|\boldsymbol{z}^{(l)}) p(\boldsymbol{z}^{(l)})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}^{(l)}|\boldsymbol{x}_o, \boldsymbol{x}_m^0)} \right] \tag{5.33}$$

Ablations. Our full model trains using MCEM and SIR and models the VAE prior using a NF. To explore the value addded by each of these components, we run a series of ablation

studies in which VAEs are trained using MCEM, SIR, or both with or without a NF prior.

Training and Hyperparameter Tuning. Each imputer has a number of hyperparameters which must be fixed prior to training. As these settings may significantly impact training dynamics and final imputation quality, we use the Asynchronous Successive Halving Algorithm (ASHA) (Li et al. 2020) to efficiently identify performant hyperparameter settings. For each model, we define a search space comprising the hyperparameters we wish to tune. ASHA explores this space by training multiple instances of the model in parallel, monitoring their performance on held-out loss minimization, eliminating weak performers, and instantiating new models with hyperparameters sampled from promising regions of the search space.

For fairness, we run ASHA for each model keeping the computational budget uniform across models. We note briefly that the search spaces for more complex models (e.g. those including NFs) are larger and therefore more challenging to explore relative to the more constrained search spaces for our baselines. At test time, we report results for each imputer using the most performant checkpoint identifed while tuning.

Data. We use publicly available datasets to train imputers. Specifically, we focus here on two-dimensional, benchmark datasets including `two moons`, `circles`, and `blobs` (Pedregosa et al. 2011b). Though these data are relatively simple, they have interesting and useful properties which make them suitable for our studies. First, they are multimodal which implies the true conditionals, $p(x_m|x_o)$, are often multimodal as well. This makes imputation challenging since good imputers will need to approximate a wide range of multimodal conditionals and avoid simplistic, low-likelihood solutions such as mean imputation. Second, sampling new observations for these datasets is fast and efficient. This allows us to easily obtain approximate samples from $p(x_m|x_o = x'_o)$ for any pair of missing and observed variables (x'_o, x'_m). Below we describe how these samples are obtained and how they are used to evaluate the quality of imputations generated by our imputers. Finally, these two-dimensional datasets

are easily visualized. This allows us to assess imputations qualitatively via visual inspection.

For each dataset we sample 6×10^4 instances retaining 5×10^4 for training, 5×10^3 for validation, and 5×10^3 for testing. To generate partially observed data, we apply synthetic missingness masks to each dataset. These masks are constructed by first randomly selecting 20% of instances, and, for each, randomly assigning one feature as missing. We ensure the missingness rate is uniform across all dataset partitions.

For each pair of missing and observed values, (x'_o, x'_m), we obtain 2.5×10^3 approximate samples from the corresponding conditional $p(x_m|x_o = x'_o)$. This is done by a simple rejection sampling scheme in which we sample candidate instances (x^c_o, x^c_m), and retain only those samples satisfying $|x^c_o - x'_o| < \delta$. We set $\delta = 10^{-3}$ for all datasets. Next, we describe how these samples are used to evaluate our imputers.

Evaluation. We are interested in evaluating how well VAEs trained using our algorithm perform as imputers relative to other models. A good imputer should generate samples of missing values which appear to be drawn from the true conditionals $p(x_m|x_o)$. Furthermore, a good imputation model should produce imputed instances which appear to be drawn from the true data distribution.

Quantitative Evaluation. Every imputation model we consider generates samples of $\boldsymbol{x}_m$ by conditioning on observations $\boldsymbol{x}_o$. Thus, an imputation model induces a family of conditional distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_m|\boldsymbol{x}_o)$, and the optimal imputation model corresponds to the family of true conditionals, $p(\boldsymbol{x}_m|\boldsymbol{x}_o)$. We evalute imputation models based on how closely their samples approximate samples from the true conditionals. This is done using the two-sample Kolmogorov-Smirnov (KS) test, which compares the empirical CDFs (ECDFs) of two sets of samples to estimate the likelihood that they were drawn from a common distribution. For each missing value in a dataset, we obtain 2.5×10^3 samples from the imputation model and the corresponding (approximate) true conditional as described above. We then apply the KS test to each set of samples and record all the KS statistics. We use boxplots to

visualize and compare the KS test statistic distributions across imputation models (lower values are better).

Qualitative Evaluation. Ideally, imputed data should appear as though it were sampled from the true data distribution, $p(\boldsymbol{x})$. As such, for each imputer, we plot imputed samples alongside samples drawn from $p(\boldsymbol{x})$ for visual comparison.

Results

Figure 5.1 summarizes our qualitative evaluation by showing fully observed data sampled from the true data distributions alongside samples of imputed data generated by each imputation model. Imputation quality varies greatly across models. MVAE often imputes near the data's center of mass (COM) where the true data distributions lack support. MIWAE also displays this problem, though less severely. Models without a NF prior also impute away from the true data support. Training with MCEM alone results in imputations at the COM and also in non-central areas intermodal regions as can be seen in the imputed `two moons` data. Training with SIR alone generates imputations in both low density intermodal and non-intermodal regions, such as the lower left and upper right corners in the `two moons` figure. Combining MCEM and SIR seems to attenuate the problems each method shows in isolation; imputations for all datasets more closely mirror the true data distribution and seem qualitatively better and slightly worse than imputations generated by MVAE and MIWAE, respectively. Adding a NF prior seems to generally improve imputation quality. When combined with MCEM, imputations near the COM are greatly reduced, but imputing in non-intermodal areas with no true support can be seen as in the `two moons` data. Combining the NF prior with SIR nearly completely eliminates imputations away from the true support and, though some artifacts remain as in the corners of the `two moons` data. Finally, our full model combining MCEM, SIR, and the NF prior imputes nearly always within the support of the true data distribution for all datasets, and thus appear qualitatively superior to all

our ablated models and baselines.

The results of our quantitative evaluation are summarized in Figure 5.2. This evaluation is designed to measure how well each imputer approximates the true conditonal distribution $p(x_m|x_o = x'_o)$ for each partially observed data point, $\boldsymbol{x} = (x'_o, x'_m)$ in our test set. To better illustrate this process, we focus on just one such data point in the first two columns of Figure 5.2. For each dataset we choose a point with missingness in x_0 and where the observed value of x_1 is near 0. This choice ensures that the true conditional, $p(x_0|x_1 = x'_1)$, contains at least two modes. The second column shows the ECDFs for samples from the true conditonal and imputers given $x'_o = x'_1$ as input. The more similar these ECDFs appear, the better the imputer. MVAE, which often imputes near the COM, produces sigmoidal ECDFs that do not resemble any of the true sample ECDFs. MIWAE does significantly better, capturing each mode in the true conditonal and weighting them appropriately. Meawhile, the ablated models show a trend similar to that seen in then qualitative evaluation: models perform better as more components are added, and the full model using MCEM, SIR, and the NF prior appears to do the best.

The two-sample KS test also measures the similarity between two distributions by comparing the ECDFs produced by their samples. For each imputer, we calculate the KS test statistic for every missing value in the test set using samples from the corresponding true conditional and the imputer. This generates a set of KS test distributions which we visualize in the third column of Figure 5.2. The two-sample KS test statistic resides within the interval $[0, 1]$, and lower values correspond to lower likelihood of rejecting the null hypothesis both sets of samples originated from a common distribution. Thus KS test distributions that skew toward 0 are considered better. For all datasets, MVAE appears to do worse than MIWAE, which performs well against most models. The ablated models show a loose trend suggesting improvments in performance as more components are added to the model. The full model appears to do best overall, though its KS test distribution for the `circles` dataset overlaps significantly with those for MIWAE and the ablated models which use the NF prior.

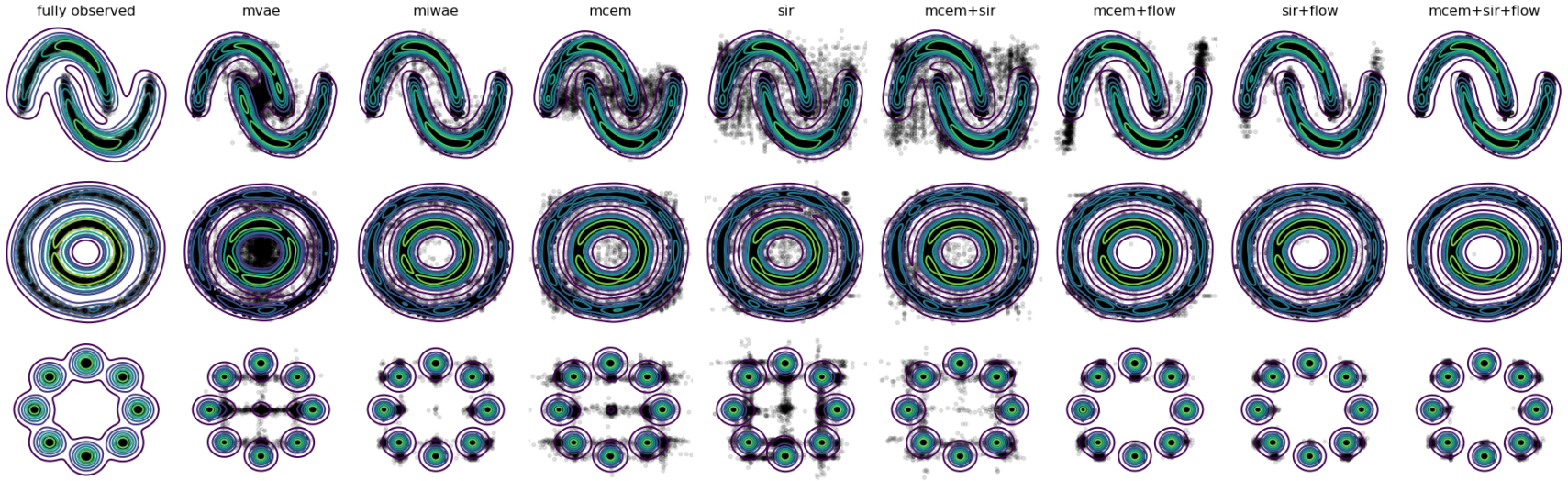

Figure 5.1: Fully observed and imputed samples for `two moons`, `circles`, and `blobs` datasets in row order. Contour lines are estimated using Gaussian kernel density estimates.

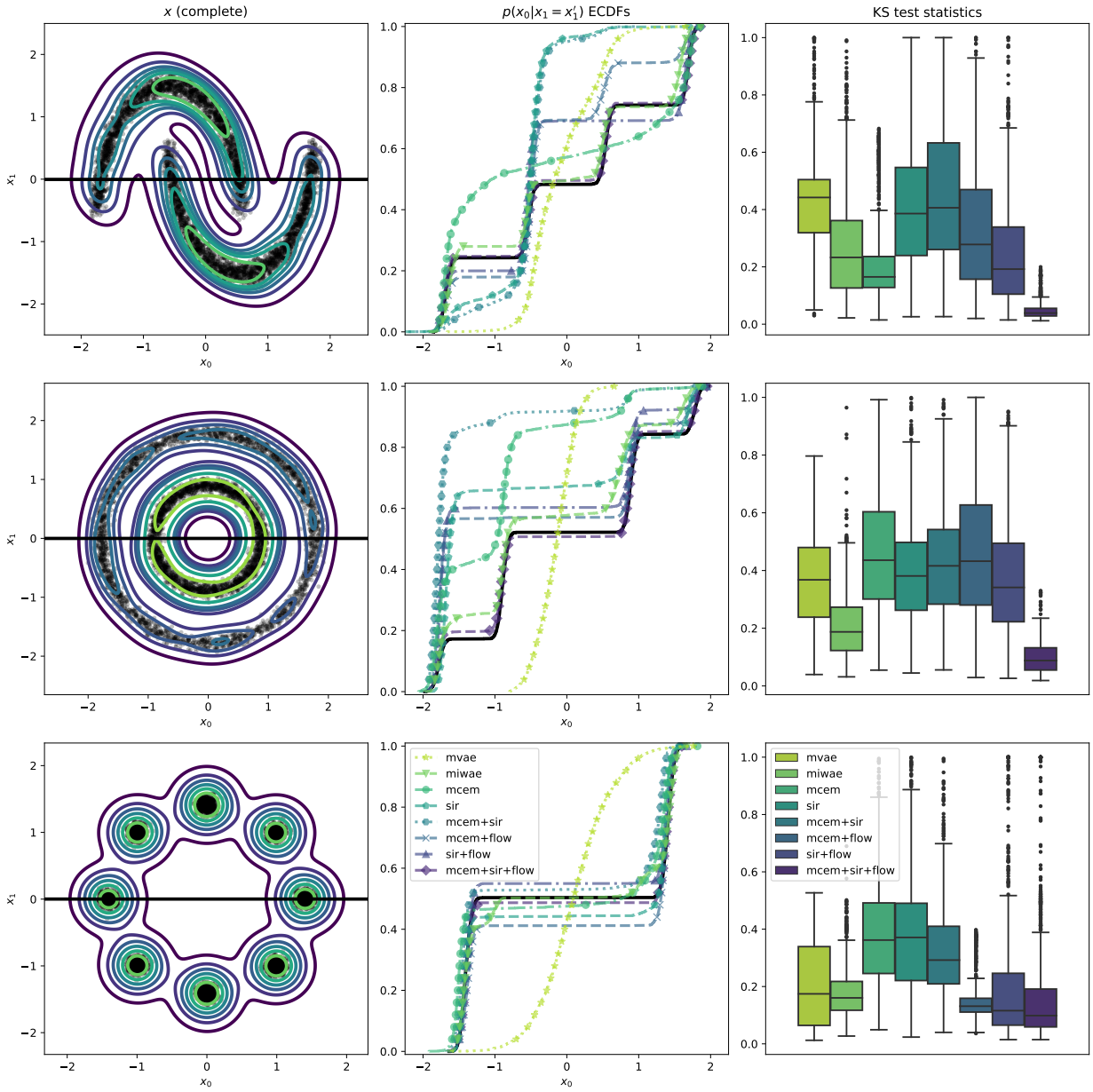

Figure 5.2: Imputation Quality. **First column**: Complete data for `two moons`, `circles`, and `blobs` datasets in row order. Contour lines are estimated using Gaussian kernel density estimates. Solid black lines are drawn at $x_1 = x_1'$. **Second column**: ECDFs for samples from the true conditional $p(x_0|x_1 = x_1')$ (solid black line) and samples generated by imputers given observed data, $x_o = x_1'$. **Third column**: Distributions of Two-sample KS test statistics generated for each imputer by comparing true conditional and imputed samples for all missing values in the test set (***lower is better***). Boxplots show medians bounded within interquartile regions (IQRs). Maximum whisker lengths are set to $1.5 \times$ IQR beyond which data are considered outliers and are shown as black points.

Discussion

In this subaim, we describe a MCEM algorithm for training VAEs on partially observed data and evaluate the resultant model on a set of low-dimensional imputation tasks. In additon, we describe two additonal components — a SIR algorithm and the incorporation of a trainable NF prior — designed to improve models trained with MCEM. Training models using all three components, yields imputers which consistently outperform baseline methods in terms of both the visual quality of imputations and their ability to approximate an ideal imputation model defined by the conditionals, $p(\boldsymbol{x}_m|\boldsymbol{x}_o)$.

Importantly, our results suggest that MCEM, SIR, and NF priors each contribute meaningfully to the final model. Across all our experiments, we observe a recurring trend in which the addition of a component results in a noticeably better imputer; no single component or subset thereof does as well all three combined. Combining SIR and MCEM may be better than using either in isolation since SIR is designed to search the latent space for regions that are better able to explain the observed data, while MCEM works to explain missing values based on the observed data by iteratively approaching the equilibrium distribution, $p_{\boldsymbol{\theta}}(\boldsymbol{x}_m|\boldsymbol{x}_o)$. Futhermore, SIR is able to sample for many regions in latent space, which can help MCEM avoid getting stuck in a single mode. Meanwhile, the NF prior adds expressivity to the VAE generative model, by adding additional parameters and removing the regularization imposed by using a standard normal prior. A multimodal latent space may be more likely when using an NF prior. This could explain why combining NF priors with just SIR seems to always improve imputation quality, but only sometimes helps when using just MCEM; the former can exploit and guide identification of multiple latent modes, while the latter is prone to sticking in single modes. Finally, combining all three results in a model that can learn a latent spaces with mutiple modes, which are identified and exploited in accordance with their ability to explain the observed data and in service of better approximating the arbitrary conditional distributions which best model the missing data.

Limitations

The generalizability of our results is limited due to our reliance upon low-dimensional data throughout our experiments. These data permit us to carry out our evaluation strategy which requires visualization, and more significantly, access to approximate samples from the true conditionals, $p(\boldsymbol{x}_m|\boldsymbol{x}_o)$. Future work should utilize alternative evaluation metrics to explore the utility of our method in higher-dimensional settings where results may better translate to general use cases for imputation. We are also limited by our use of complete data and synthetic missingness. Again, access to complete data allowed us to move forward with our evaluation strategy, but this also required that we introduce our own scheme for generating partially observed data. Since we do not condition on either the observed or missing values when deciding which values go misssing, our data are misssing completely at random (MCAR). This is the simplest kind of missingness, and it also it's use justifies training models to maximize the observed data likelihood (or a lower bound on it) (Rubin 1976). Unfortunately, MCAR missingness is rarely, if ever, found in real data, which are more likely go missing conditonal on the observed data (i.e. missing at random (MAR)), or, more likely still, the missing values themselves (ie. missing not at random (MNAR)). Future work evaluationg imputation quality will likely still need to use complete datasets, but synthetic MAR and MNAR missignness could be applied. Aside from the type of missingness used, we also only experiment with missignness in 20% of the data. Real data may have more or less severe missigneess, and future work should experiment with a range of missigness rates.

5.2 Aim 3B. Develop a semi-supervised VAE which can be trained on partially observed data using MCEM; use it for disease phenotyping on clinical data with inherent missingness.

Background

Classification problems are defined for datasets whose instances comprise two disjoint sets — features, $\boldsymbol{x}_c$ and labels, $\boldsymbol{y}_c$ — where the latter are constrained to discrete values (e.g. binary, categorical). Typically, the goal is to obtain a model which conditions on features to estimate the likelihood of the labels, $p(\boldsymbol{y}_c|\boldsymbol{x}_c)$. Alternatively, one may choose to model the features and labels jointly, $p(\boldsymbol{x}_c, \boldsymbol{y}_c)$. Here, we adopt the latter approach.

VAEs handle heterogeneous data naturally. Since the likelihood function specified by the decoder assumes conditional independence among the variables $x_j \in \boldsymbol{x}$ (i.e $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) = \prod_j p_{\boldsymbol{\theta}}(x_j|\boldsymbol{z})$) one may simply model each conditional factor using a suitable likelihood (e.g. normal for continuous variables, Bernoulli for binary variables). In the present subaim, we combine this feature with MCEM to solve classification problems where both $\boldsymbol{x}_c$ and $\boldsymbol{y}_c$ contain missing values. As such, our VAEs achieve semi-supervised learning (missingness in $\boldsymbol{y}_c$) while requiring only partially observed features (missingness in $\boldsymbol{x}_c$).

As in the previous subaim, we first illustrate the utility of our approach using low-dimensional data. We then experiment with semi-supervised phenotyping: we train VAEs on partially observed clinical data and use the resultant models to predict a set of labels encoding multiple, binary patient states.

Research Question

Does a VAE trained on partially labeled, partially observed data yield an effective disease phenotyping model?

Methods

Models. For VAEs trained using MCEM, we use SIR and employ architectures containing a normalizing flow prior (MAF). We compare against VAEs trained using the MIWAE objective. This choice is guided by our experiments in the previous subaim, where these configurations led to the best performance among MCEM-optimized VAEs and baselines, respectively.

As detailed below, our datasets comprise continuous-valued features and binary labels. Thus, for all VAEs, we specify a heterogeneous likelihood function in the decoder which places conditionally independent normal and Bernoulli distributions over the feature and label variables, respectively. Aside from this modification, the models and their training procedures remain unchanged.

Training and Hyperparameter Tuning. As described in the previous subaim, we use ASHA to tune hyperparameters for each combination of model and dataset. We hold the computational budget fixed across all models trained on a given dataset to ensure fairness at test time. For each model, we report results only for the best identified checkpoint.

Data. Our first dataset is `two moons` which serves to define a low-dimensional classification problem. Here the features correspond to the two continuous values specifying the location of each point in the plane; a single binary label specifies which moon (top or bottom) a point in the plane belongs to (see Figure 5.3). As in the previous subaim, we sampled 6×10^4 instances and randomly partition out 5×10^4 for training, 5×10^3 for validation, and 5×10^3 for testing. We apply synthetic missingness masks constructed by randomly eliminating 20% of observations while ensuring each instance retains at least one observed value.

Our second dataset is derived from the publicly available Women in Data Science (WiDS) Datathon 2020: ICU Mortality Prediction data (Lee et al. 2020). These data contain a heterogenous mixture of partially observed clinical variables recorded during 130,000 ICU

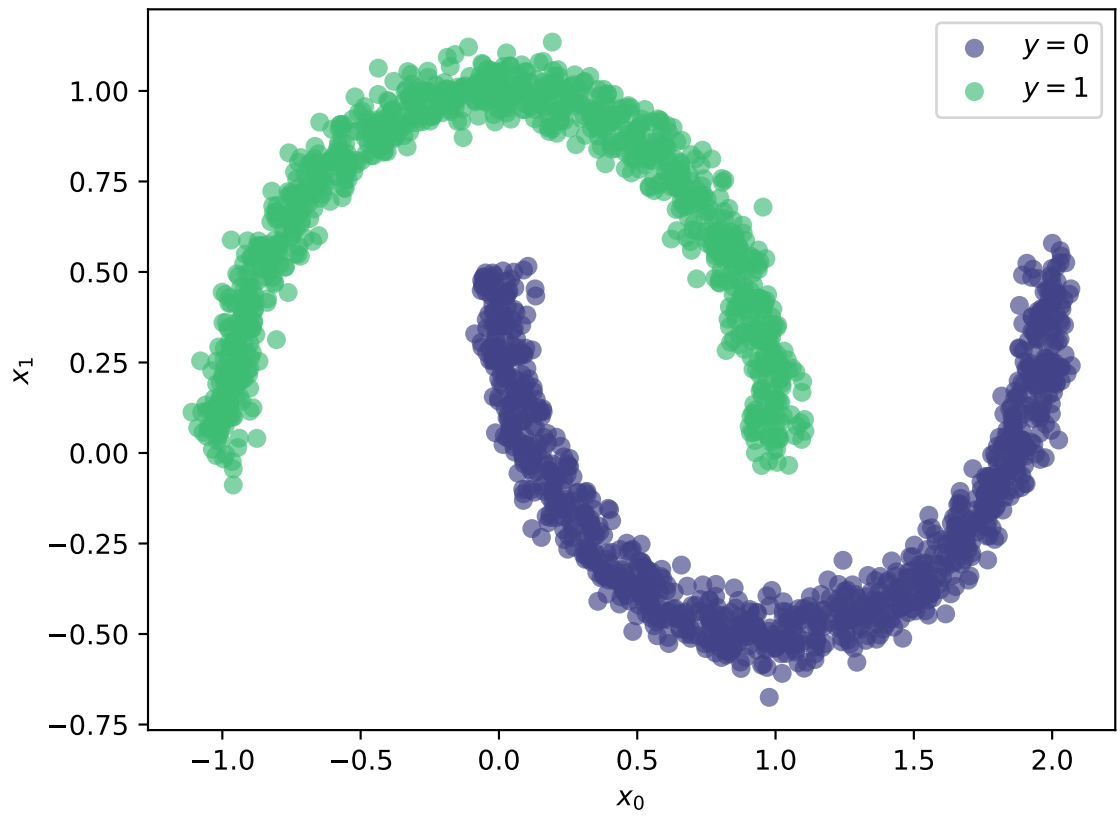

Figure 5.3: Labeled `two moons` data. Features correspond to two-dimensional coordinates, x_0 and x_1. y encodes the binary label specifying the "moon" each point belongs too: $y = 0$ and $y = 1$ for the bottom and top moon, respectively.

stays at hospitals in Argentina, Australia, New Zealand, Sri Lanka, and the United States. Though these data were curated primarily to train in-patient mortality prediction models, they also contain binary values encoding a variety of patient states (e.g. is intubated, is diabetic). For our label set, we consider only those patient states with low missingness and high prevalance; these are `ventilated` (prevalence: 0.325), `intubated` (prevalence: 0.151), `diabetes mellitus` (prevalence: 0.225), and `hospital death` (prevalence: 1.0). For the feature set, we include only the continuous-valued variables. See Table 5.1 for more details on the selected variables and their intrinsic missingness rates.

Our evaluation strategy (discussed below) requires both predictions for missing labels and access to their ground truth values. However, unlike our low-dimensional dataset, the WiDS data are *inherently* partially observed. Thus, for some hospital stays, ground truth values are not available for each label. To remedy this conflict, we only consider those hospital stays with a complete label set. This reduces the total number of hospital stays to 90,998 which we then parition (80% train, 10% validation, 10% test). We then apply 20% synthetic missingness to the labels. For the features, missingness is inherent; no synthetic missingness is applied. See Table 5.1 for a list of variables included in our preprocessed dataset and their corresponding missingness rates.

Evaluation. We evaluate models based on their ability to predict the true values of missing labels in held-out data. To do so, we report the areas under the ROC and Precision-Recall curves (AUROC and AUPRC, respectively). In the multi-label setting, we report these metrics per label, as well as the micro- and macro-averages over all labels.

Results

Table 5.2 displays results for our low-dimensional experiments with the `two-moons` data. Recall, the task here is to correctly predict which moon (top or bottom) an instance belongs to when the label is missing and the features are partially observed. MCEM-SIR outperforms MIWAE in both AUROC and AUPRC.

	AUROC	AUPRC
MIWAE	0.766	0.735
MCEM-SIR	**0.881**	**0.860**

Table 5.2: Classification metrics for held-out `two-moons` data.

Results for our phenotyping experiments are shown in Table 5.3. In this setting, MCEM-SIR and MIWAE demonstrate similar performance in both metrics across all labels and their

Type	Variable	Description	Unit	MR	1hr MR	24hr MR	APACHE MR
	age	Age	Years	0.039			
	height	Height	cm	0.015			
	weight	Weight	kg	0.030			
	bmi	Body mass index	kg/m^3	0.037			
	pre_icu_los_days	Length of stay	Days	0.000			
	heartrate	Heart rate	Beats/min		0.030	0.002	0.002
	resprate	Respiratory rate	Breaths/min		0.047	0.004	0.006
	temp	Core temperature	°C		0.236	0.025	0.037
	spo2	Peripheral oxygen saturation	%		0.045	0.004	
	arterial_pco2	Arterial partial pressure of carbon dioxide	mmHg		0.828	0.646	0.771
	arterial_ph	Arterial pH	None		0.833	0.655	0.771
	arterial_po2	Arterial partial pressure of oxygen	mmHg		0.828	0.646	0.771
	pao2fio2ratio	Fraction of inspired oxygen	None		0.874	0.720	0.771
	urineoutput_apache	Total urine output for the first 24 hours	mL				0.531
	diasbp_invasive	Diastolic blood pressure, invasively measured	mmHg		0.817	0.741	
	diasbp_noninvasive	Diastolic blood pressure, non-invasively measured	mmHg		0.080	0.011	
	diasbp	Diastolic blood pressure, either non-invasively or invasively measured	mmHg		0.039	0.002	
	sysbp_invasive	Systolic blood pressure, invasively measured	mmHg		0.817	0.741	
Features	sysbp_noninvasive	Systolic blood pressure, non-invasively measured	mmHg		0.080	0.011	
	sysbp	Systolic blood pressure, either non-invasively or invasively measured	mmHg		0.039	0.002	
	mbp_invasive	Mean blood pressure, invasively measured	mmHg		0.816	0.739	
	mbp_noninvasive	Mean blood pressure, non-invasively measured	mmHg		0.099	0.016	
	mbp	Mean blood pressure, either non-invasively or invasively measured	mmHg		0.050	0.002	0.003
	sodium	Sodium concentration in serum or plasma	mmol/L		0.792	0.111	0.197
	potassium	Potassium concentration in serum or plasma	mmol/L		0.786	0.104	
	hco3	Bicarbonate concentration in serum or plasma	mmol/L		0.830	0.164	
	calcium	Calcium concentration in serum	mmol/L		0.827	0.142	
	bun	Blood urea nitrogen concentration in serum or plasma	mmol/L		0.819	0.114	0.204
	glucose	Glucose concentration in serum or plasma	mmol/L		0.573	0.063	0.113
	creatinine	Creatinine concentration in serum or plasma	μmol/L		0.817	0.111	0.199
	wbc	White blood cell count	10^9/L		0.828	0.143	0.234
	hemaglobin	Hemoglobin concentration	g/dL		0.797	0.132	
	hematocrit	Volume proportion of red blood cells blood	None		0.800	0.127	0.211
	platelets	Platelet count	10^9/L		0.825	0.146	
	bilirubin	Bilirubin concentration in serum or plasma	μmol/L		0.923	0.585	0.631
	albumin	Albumin concentration in serum	g/L		0.914	0.535	0.590
	inr	International normalized ratio	μmol/L		0.632	0.632	
	lactate	Lactate concentration in serum or plasma	mmol/L		0.920	0.746	
	ventilated	Whether the patient was invasively ventilated	None	0.000			
Labels	intubated	Whether the patient was intubated	None	0.000			
	diabetes_mellitus	Whether the patient has been diagnosed with diabetes	None	0.000			
	hospital_death	Whether the patient died during this hospitalization	None	0.000			

Table 5.1: Features and labels extracted from WiDS Datathon 2020: ICU Mortality Prediction dataset and their intrinic missingness rates. Many variables have repeated measurements which are represented in the dataset by the minimum and maximum values observed within the first hour and first 24 hours. Some variables also contribute to the Acute Physiology and Chronic Health Evaluation III (APACHE III) severity-of-disease classification system, and the values used in the score are represented by additional variables in the dataset. Variables measured only once (e.g. age) have one missingness rate shown in column *MR*. Variables with repeated measurements (e.g. sodium) have multiple missingness rates: missingness for the minimum and maximum values within the first hour and first 24 hours shown in columns *1hr MR* and *24hr MR*, respectively. Missingness rates for variables used in APACHE III (e.g. temp) are shown in column *APACHE MR*. Note that for any variable with repeated measurements the missingness rates for the minimum and maximum values are identical. Thus, we represent both variables with a single variable name.

averages, with neither method being consistently superior.

	AUROC		AUPRC	
Label	MIWAE	MCEM-SIR	MIWAE	MCEM-SIR
`ventilated`	**0.846**	0.842	**0.743**	0.735
`intubated`	0.827	**0.844**	**0.440**	0.438
`diabetes mellitus`	0.795	**0.805**	0.525	**0.539**
`hospital death`	**0.834**	0.811	**0.374**	0.352
macro average	0.825	**0.826**	0.600	**0.602**
micro average	0.848	**0.850**	**0.521**	0.516

Table 5.3: Classification metrics for held-out WiDS data.

Discussion

This subaim explores the use of MCEM for training VAEs on partially observed, partially labeled data. In a low-dimensional setting, our results suggest that MCEM (in combination with SIR and an NF prior) yields a VAE which is better able to perform as a classifier relative to a MIWAE baseline. Recall in the previous subaim that, relative to MIWAE, MCEM was better able to approximate true conditionals defined by the `two-moons` data distribution. This is likely a driving factor behind MCEM's superior performance in the present context as well. Consider the true conditional for the label, y_c, when the features are both observed, $p(y_c|\boldsymbol{x}_c)$. If the model well approximates this conditional, then it should be a strong classifier given that the moon's are disjoint in the plane. Meanwhile, if only one of the features is observed, approximating the true conditional should still result in good classifications in areas where there is no overlap in the moons' projections along the observed dimension. Thus, at least in a low-dimensional settings, MCEM does appear to be a viable strategy for handling classification problems when both the features and labels contain missing values.

We also explore the use of MCEM for clinical phenotyping. This setting is significantly

more complex: the data are higher dimensional, the task is multi-label, and the features contain inherent missingness (i.e. we do not know the true missing values, we do not control the missingness mechanism, and it is likely not MCAR). Our results show that, though MCEM achieves relatively high AUROCs and AUPRCs for these data, it does not clearly outperform MIWAE. There are several possible explanations for this outcome. Since the data are high-dimensional, and any instance may have multiple missing values, each model is tasked with estimating many multivariate conditional distributions. It is possible that, under either model, we simply lack sufficient observations to estimate each conditional faithfully, and thus both perform similarly. Alternatively, the structure of the true data distribution may be to blame. For example, it may be the case that many true conditionals for the binary labels do not clearly distinguish between the two states. In this scenario, both models would perform similarly since they are both attempting to model many true conditionals which are ill-suited for prediction.

Limitations

The work described in this subaim has several limitations. Similar to our previous subaim, we rely on complete data in our labels to carry out our evaluation. This allows us to use standard supervised learning metrics to evaluate our models, but it also requires that we design our own missingness mechanism for the labels. This missingness is MCAR and is applied to 20% of the observations. Future work should explore more complex missingness mechanisms and a wider distribution of missingness rates. However, for our phenotyping task, it should be noted that this limitation is somewhat attenuated by the inherent missingness in the features, which is likely not MCAR but rather MNAR. Our clinical dataset also introduces some limitations. It was developed for use in a machine learning competetion and is thus highly curated. Data derived from an average hospital's data stores is likely to be much noisier. Furthermose, though these data permit us to define a multi-label phenotyping problem, the dataset was originally developed specifically for mortality prediction.

This may have influenced which patient visits were included in the dataset, making it less represenative of clinical datasets in general.

Chapter 6

Conclusion

Determining if a patient has a particular condition can be a difficult task even for an experienced clinician with access to the patient and their full clinical history. It should be of no surprise then that accomplishing the same task for a population of patients using only a subset of their clincal records can be very challenging. Nevertheless, such phenotyping problems are routinely encountered when using clincal data for research purposes, and represent a persistent bottleneck slowing the extraction of clinically meaningful insights from growing clincal data stores. The work detailed in this dissertation has sought to loosen this constraint by introducing methods for learning phenotypes directly from data, particularly when the data are not fully observed and fully labeled, as is common in the clincal domain.

Aim 1 of this dissertation implements a previously described unsupervised method, Multi-Channel Mixed Membership Models (MC3M), and applys it to infer phenotypes from partially observed data. We train MC3M on survey questionairres developed to probe the psychosocial and behavioral constructs underlying the health of individuals. MC3M phenotypes identified subgroups within the surveyed population, which were characterized by both mutable and immutable characteristics (i.e. mediators and moderators) recorded in the data and made prominent by the phentoypes' structure. Modeling both sets of factors within a phenotyping algorithm is a signficant contribution of this work; though immutable factors are fixed, they provide essential context which may help to define and distinguish among population subgroups. Meanwhile, prominent mutable factors which emerge within subgroups may serve as targets for behavioral-health interventions. As such, our work represents a step toward the goal of intervention tailoring, which would aim to improve the

health of a population by designing interventions suited to the subpopulations it contains. We took a further step in this direction by isolating a subset of phenotypes found to have a significant association with elevated weight status, an import, mutable risk factor for chronic disease. Interrogating the structure of these phenotypes revealed salient characteristics which could be useful in tailoring weight loss interventions for specific subgroups. Thus, this work provides a proof of concept that unsupervised phenotyping can identify subpopulation and surface defining characteristics relevant to intervention tailoring.

In Aim 2 we introduce Semi-Supervised Mixed Membership Models (SS3M), a phenotyping method which builds upon MC3M by modeling a set of partially observed labels. SS3M seeks to strike a balance between unsupervised phenotyping, which can be difficult to evaluate, and manual chart review which is often prohibitively expensive. By assuming access to a small amount of labeled data, SS3M reduces phenotyping to a multi-label prediction task and minimizes the required effort spent on label generation. Our first set of experiments evaluate SS3M using simulated data, and demonstrate the model successfully learns both the structure and identity of a known, ground truth set of phenotypes. Notably, SS3M accomplishes this task with only 5% of the true labels made available for training. Next, we focus on learning phenotypes from actual clinical data. In this setting, SS3M outperformed MC3M on a set of evaluations designed to measure phenotype quality and interpretability from the perspecitive of clinical experts. Importantly, our surveyed clinical experts found that a large majority of SS3M phenotypes captured the clincial characteristics of the conditions specified by their corresponding labels. In addtion, SS3M demonstrates competitive performance relative to supervised baselines on a set of disease prediction tasks. In summary, SS3M is shown to be an effective model for learning interpretable phenotypes from partially labeled clincal data.

Aim 3, our final aim, describes an algorithm for training Variational Autoencoders (VAEs), on partially observed data. Our approach takes a previously described algorithm for heldout data imputation, and incorporates it within a Monte Carlo Expectation-Maximization

(MCEM) scheme optimizing a lower bound on the marginal likelihood of the observed data. This algorithm allows VAEs to to train on data containing missing values, but also has some theoretical limitations which try to overcome by 1) adding a Sampling Importance Resampling (SIR) step to MCEM, and 2) increasing the modeling capacity of VAEs by using a normalizing flow (NF) as the latent space prior. Experiments with low-dimensional datasets show our approach results in imputations which consistently lay within the support of the true data distribution — a result that baselines struggle or fail to replicate. Furthermore, among all tested imputers, VAEs trained with our method are found to most closely approximate the optimal imputation model represented by the true conditionals, $p(x_m|x_o)$. Though our results on imputation are limited to low-dimensional data, our method shows promise, and we suspect similarly strong performance will be seen when it is applied to more complex datasets. We also experiment with MCEM in the context of semi-supervised learning. In a low-dimensional settings, VAEs trained with MCEM outperform a strong baseline on label prediction. We also apply this strategy for multi-label phenotyping, where MCEM-trained VAEs perform similar to baseline. Nevertheless, MCEM shows promise in this space and future work will aim to improve performance.

Bibliography

Agarwal, Vibhu et al. (Nov. 1, 2016). "Learning statistical models of phenotypes using noisy labeled training data". In: *Journal of the American Medical Informatics Association* 23.6, pp. 1166–1173.

Ahuja, Yuri et al. (Aug. 1, 2020). "sureLDA: A multidisease automated phenotyping method for the electronic health record". In: *Journal of the American Medical Informatics Association* 27.8, pp. 1235–1243.

Amw, LeBrón et al. (2015). "Socioeconomic position, and blood pressure in a multi-ethnic urban community". In: *Ethn Dis* 25, pp. 24–30.

Aronson, A. R. (2001). "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." In: *Proceedings of the AMIA Symposium*, pp. 17–21.

Aronson, Alan R and François-Michel Lang (May 1, 2010). "An overview of MetaMap: historical perspective and recent advances". In: *Journal of the American Medical Informatics Association* 17.3, pp. 229–236.

Banda, Juan M. et al. (July 26, 2017). "Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network". In: *AMIA Summits on Translational Science Proceedings* 2017, pp. 48–57.

Banda, Juan M. et al. (July 20, 2018). "Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models". In: *Annual Review of Biomedical Data Science* 1.1. Publisher: Annual Reviews, pp. 53–68.

Bayley, K. Bruce et al. (2013). "Challenges in Using Electronic Health Record Data for CER: Experience of 4 Learning Organizations and Solutions Applied". In: *Medical Care* 51. Publisher: Lippincott Williams & Wilkins, S80–S86.

Bejan, Cosmin A. et al. (2013). "On-time clinical phenotype prediction based on narrative reports." In: *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2013, pp. 103–110.

Bergquist, Savannah L et al. (2017). "Classifying lung cancer severity with ensemble machine learning in health care claims data". In: *Proceedings of machine learning research* 68, p. 25.

Bhattacharya M. et al. (2017). "Identifying ventricular arrhythmia cases and their predictors by applying machine learning methods to electronic health records (EHR) of hypertrophic cardiomyopathy (HCM) patients". In: *Circulation* 136 ((Bhattacharya M.; Shatkay H.) Computer and Information Sciences, Univ of Delaware, Newark, DE, United States). Number: (Bhattacharya M.; Shatkay H.) Computer and Information Sciences, Univ of Delaware, Newark, DE, United States.

Bishop, Christopher (Jan. 2006). *Pattern Recognition and Machine Learning*. Springer.

Blecker, Saul et al. (Sept. 5, 2017). "Early Identification of Patients With Acute Decompensated Heart Failure." In: *Journal of cardiac failure*.

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe (Apr. 3, 2017). "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2017.1285773, pp. 859–877.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.

Boden-Albala, B. et al. (2011). "Perception of recurrent stroke risk among black, white and Hispanic ischemic stroke and transient ischemic attack survivors: the SWIFT study". In: *Neuroepidemiology* 37, pp. 83–87.

Booth, J. M. and Jonassaint Cr (2016). "The Role of Disadvantaged Neighborhood Environments in the Association of John Henryism With Hypertension and Obesity". In: *Psychosom Med* 78, pp. 552–561.

Bouhlal, S. et al. (2017). "Identifying eating behavior phenotypes and their correlates: A novel direction toward improving weight management interventions". In: *Appetite* 111, pp. 142–150.

Boutelle, K. N. et al. (2014). "Overeating phenotypes in overweight and obese children". In: *Appetite* 76, pp. 95–100.

Bryan, C. J., E. Tipton, and Yeager Ds (2021). "Behavioural science is unlikely to change the world without a heterogeneity revolution". In: *Nat Hum Behav* 5, pp. 980–989.

Burgermaster, M. et al. (2017). "Testing an Integrated Model of Program Implementation: the Food, Health Choices School-Based Childhood Obesity Prevention Intervention Process Evaluation". In: *Prev Sci* 18, pp. 71–82.

Burgermaster, M. et al. (2018). "Behavior change is not one size fits all: psychosocial phenotypes of childhood obesity prevention intervention participants". In: *Transl Behav Med* 8, pp. 799–807.

Buysse, D. J., L. Yu, and et al. Moul D. E. (2010). "Development and Validation of Patient-Reported Outcome Measures for Sleep Disturbance and Sleep-Related Impairments". In: *Sleep* 33, pp. 781–792.

Califf, Robert M. (May 1, 2014). "The Patient-Centered Outcomes Research Network: A National Infrastructure for Comparative Effectiveness Research". In: *North Carolina Medical Journal* 75.3. Publisher: North Carolina Medical Journal Section: Policy Forum, pp. 204–210.

Calugi, S. and R. Dalle Grave (2020). "Psychological features in obesity: A network analysis". In: *Int J Eat Disord* 53, pp. 248–255.

Cantor, D. et al. (2009). *Health Information National Trends Survey 2007*. Final Rep Natl Cancer Inst.

Carroll, Robert J., Anne E. Eyler, and Joshua C. Denny (2011). "Naive Electronic Health Record phenotype identification for Rheumatoid arthritis." In: *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2011, pp. 189–196.

Centers for Disease Control and Prevention Behavioral Risk Factor Surveillance System (2010).

Cerna A.E.U. et al. (2017). "Data driven phenotyping of patients with heart failure using a deep-learning cluster representation of echocardiographic and electronic health record data". In: *Circulation* 136 ((Cerna A.E.U.; Wehner G.; Hartzel D.N.; Haggerty C.; Fornwalt B.) Imaging Science and Innovation, Geisinger, Danville, PA, United States). Number: (Cerna A.E.U.; Wehner G.; Hartzel D.N.; Haggerty C.; Fornwalt B.) Imaging Science and Innovation, Geisinger, Danville, PA, United States.

Chaganti, Shikha et al. (Feb. 11, 2017). "Phenotype Analysis of Early Risk Factors from Electronic Medical Records Improves Image-Derived Diagnostic Classifiers for Optic Nerve Pathology." In: *Proceedings of SPIE–the International Society for Optical Engineering* 10138.

Chapelle, Olivier, Bernhard Schölkopf, and Alexander Zien, eds. (Sept. 22, 2006). *Semi-Supervised Learning*. Red. by Francis Bach. Adaptive Computation and Machine Learning series. Cambridge, MA, USA: MIT Press. 528 pp. ISBN: 978-0-262-03358-9.

Che, C. et al. (June 9, 2017). "An RNN Architecture with Dynamic Temporal Matching for Personalized Predictions of Parkinson's Disease". In: *Proceedings of the 2017 SIAM International Conference on Data Mining*. 0 vols. Proceedings. Society for Industrial and Applied Mathematics, pp. 198–206.

Che, Zhengping et al. (Aug. 10, 2015a). "Deep Computational Phenotyping". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15. New York, NY, USA: Association for Computing Machinery, pp. 507–516. ISBN: 978-1-4503-3664-2.

Che, Zhengping et al. (Dec. 11, 2015b). "Distilling Knowledge from Deep Networks with Applications to Healthcare Domain". In: *arXiv:1512.03542 [cs, stat]*. arXiv: 1512.03542.

Chew, L. D., J. M. Griffin, and et al. Partin M. R. (2008). "Validation of Screening Questions for Limited Health Literacy in a Large VA Outpatient Population". In: *J Gen Intern Med* 23, pp. 561–566.

Chubachi, Shotaro et al. (Aug. 2016). "Identification of five clusters of comorbidities in a longitudinal Japanese chronic obstructive pulmonary disease cohort." In: *Respiratory medicine* 117, pp. 272–279.

Churchill, Frederick B. (1974). "William Johannsen and the Genotype Concept". In: *Journal of the History of Biology* 7.1. Publisher: Springer, pp. 5–30.

Coghlan, Simon and Simon D'Alfonso (Dec. 1, 2021). "Digital Phenotyping: an Epistemic and Methodological Analysis". In: *Philosophy & Technology* 34.4, pp. 1905–1928.

Cohen, S., T. Kamarck, and R. : A. Mermelstein (1983). "Global Measure of Perceived Stress". In: *J Health Soc Behav* 24, pp. 385–396.

Collins, F. S. and H. : A. Varmus (2015). "new initiative on precision medicine". In: *N Engl J Med* 372, pp. 793–795.

Dawkins, Richard (1982). *The extended phenotype : the gene as the unit of selection*. San Francisco: Freeman. ISBN: 978-0-7167-1358-6.

Degner, L. F., J. A. Sloan, and P. Venkatesh (1997). "The control preferences scale". In: *Can J Nurs Res Arch* 29.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1. Publisher: [Royal Statistical Society, Wiley], pp. 1–38.

Denny, Joshua C. et al. (Oct. 7, 2011). "Variants Near FOXE1 Are Associated with Hypothyroidism and Other Thyroid Conditions: Using Electronic Medical Records for Genome- and Phenome-wide Studies". In: *The American Journal of Human Genetics* 89.4, pp. 529–542.

Ding, Daisy Yi et al. (2019). "The Effectiveness of Multitask Learning for Phenotyping with Electronic Health Records Data". In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 24, pp. 18–29.

Dinh, Laurent, Jascha Sohl-Dickstein, and Samy Bengio (July 21, 2022). "Density estimation using Real NVP". In: International Conference on Learning Representations.

Dligach, Dmitriy, Timothy Miller, and Guergana K. Savova (Nov. 5, 2015). "Semi-supervised Learning for Phenotyping Tasks". In: *AMIA Annual Symposium Proceedings* 2015, pp. 502–511.

Donahue, K. et al. (2004). Identifying supports and barriers to physical activity in patients at risk for diabetes.

Doshi-Velez, Finale, Yaorong Ge, and Isaac Kohane (Jan. 2014). "Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis." In: *Pediatrics* 133.1. Number: 1, e54–63.

Dugger, S. A., A. Platt, and Goldstein Db (2018). "Drug development in the era of precision medicine". In: *Nat Rev Drug Discov* 17, p. 183.

Eaton, W. W. et al. (2004). "Center for Epidemiologic Studies Depression Scale: Review and Revision (CESD and CESD-R)". In: *In: The use of psychological testing for treatment planning and outcomes assessment: Instruments for adults* 3.3, pp. 363–377.

Essay, Patrick, Jarrod Mosier, and Vignesh Subbian (Apr. 15, 2020). "Rule-Based Cohort Definitions for Acute Respiratory Failure: Electronic Phenotyping Algorithm". In: *JMIR Medical Informatics* 8.4. Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada, e18402.

Esteban, Santiago et al. (Dec. 2017a). "A rule-based electronic phenotyping algorithm for detecting clinically relevant cardiovascular disease cases". In: *BMC Research Notes* 10.1. Number: 1 Publisher: BioMed Central, pp. 1–7.

Esteban, Santiago et al. (2017b). "Development and validation of various phenotyping algorithms for Diabetes Mellitus using data from electronic health records". In: *Computer methods and programs in biomedicine* 152, pp. 53–70.

Fan, Jin et al. (Dec. 1, 2013). "Billing code algorithms to identify cases of peripheral artery disease from administrative data". In: *Journal of the American Medical Informatics Association* 20 (e2), e349–e354.

Ferté, Thomas et al. (2021). "Automatic phenotyping of electronical health record: PheVis algorithm". In: *Journal of Biomedical Informatics* 117, p. 103746.

Freimer, Nelson and Chiara Sabatti (May 2003). "The Human Phenome Project". In: *Nature Genetics* 34.1. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group, pp. 15–21.

Fuentes, S., R. Brondeel, and et al. Franco M. (2019). "Psycho-social factors related to obesity and their associations with socioeconomic characteristics: the RECORD study". In: *Eat Weight Disord-Stud Anorex Bulim Obes*, pp. 1–11.

Furbank, Robert T. and Mark Tester (Dec. 1, 2011). "Phenomics – technologies to relieve the phenotyping bottleneck". In: *Trends in Plant Science* 16.12, pp. 635–644.

Garcia, R. M., R. B. Taylor, and Lawton Ba (2007). "Impacts of Violent Crime and Neighborhood Structure on Trusting Your Neighbors". In: *Justice Q* 24, pp. 679–704.

Garla, Vijay, Caroline Taylor, and Cynthia Brandt (Oct. 1, 2013). "Semi-supervised clinical text classification with Laplacian SVMs: An application to cancer case management". In: *Journal of Biomedical Informatics* 46.5, pp. 869–875.

Gehrmann, Sebastian et al. (Feb. 15, 2018). "Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives". In: *PLOS ONE* 13.2. Publisher: Public Library of Science, e0192360.

Gelman, Andrew et al. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Geraci, Joseph et al. (Aug. 1, 2017). "Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression". In: *Evidence-Based*

Mental Health 20.3. Publisher: Royal College of Psychiatrists Section: Original article, pp. 83–87.

Geva, Alon et al. (Sept. 2017). "A Computable Phenotype Improves Cohort Ascertainment in a Pediatric Pulmonary Hypertension Registry." In: *The Journal of pediatrics* 188, 224–231.e5.

Glanz, K., B. K. Rimer, and Viswanath K. : Health Behavior: Theory (2015). *Research, and Practice*. John Sons: Wiley.

Glicksberg, Benjamin S. et al. (2018). "Automated disease cohort selection using word embeddings from Electronic Health Records." In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 23, pp. 145–156.

Gligorijevic, Djordje, Jelena Stojanovic, and Zoran Obradovic (Dec. 1, 2016). "Disease types discovery from a large database of inpatient records: A sepsis study". In: *Methods*. Big Data Bioinformatics 111, pp. 45–55.

Griffiths, Thomas L. and Mark Steyvers (Apr. 6, 2004). "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101 (suppl 1). Publisher: National Academy of Sciences Section: Colloquium, pp. 5228–5235.

Gustafson, Erin et al. (Aug. 2017). "A Machine Learning Algorithm for Identifying Atopic Dermatitis in Adults from Electronic Health Records." In: *IEEE International Conference on Healthcare Informatics. IEEE International Conference on Healthcare Informatics* 2017, pp. 83–90.

Hahn, E. A., R. F. DeVellis, and et al. Bode R. K. (2010). "Measuring social health in the patient-reported outcomes measurement information system (PROMIS): item bank development and testing". In: *Qual Life Res* 19, pp. 1035–1044.

Halpern, Yoni, Steven Horng, and David Sontag (Dec. 10, 2016). "Clinical Tagging with Joint Probabilistic Models". In: *Proceedings of the 1st Machine Learning for Healthcare Conference*. Machine Learning for Healthcare Conference. ISSN: 1938-7228. PMLR, pp. 209–225.

Halpern, Yoni et al. (Nov. 14, 2014). "Using Anchors to Estimate Clinical State without Labeled Data". In: *AMIA Annual Symposium Proceedings* 2014, pp. 606–615.

Halpern, Yoni et al. (July 1, 2016). "Electronic medical record phenotyping using the anchor and learn framework". In: *Journal of the American Medical Informatics Association* 23.4, pp. 731–740.

Harutyunyan, Hrayr et al. (June 17, 2019). "Multitask learning and benchmarking with clinical time series data". In: *Scientific Data* 6.1. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Databases;Disease-free survival;Machine learning Subject_term_id: databases;disease-free-survival;machine-learning, p. 96.

Health Statistics, National Center for et al. (2014). *National Health and Nutrition Examination Survey. Questionnaires, datasets, and related documentation.*

Heintzelman N.H. et al. (2013). "Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text". In: *Journal of the American Medical Informatics Association* 20.5. Number: 5, pp. 898–905.

Henderson, J. et al. (Aug. 2017). "Granite: Diversified, Sparse Tensor Factorization for Electronic Health Record-Based Phenotyping". In: *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. 2017 IEEE International Conference on Healthcare Informatics (ICHI), pp. 214–223.

Henderson, Jette et al. (Dec. 5, 2018). "Phenotyping through Semi-Supervised Tensor Factorization (PSST)". In: *AMIA Annual Symposium Proceedings* 2018, pp. 564–573.

Hickey, K. T., S. Bakken, and et al. Byrne M. W. (2019). "Precision health: Advancing symptom and self-management science". In: *Nurs Outlook* 67, pp. 462–475.

Ho, Joyce C., Joydeep Ghosh, and Jimeng Sun (2014a). "Marble: High-throughput Phenotyping from Electronic Health Records via Sparse Nonnegative Tensor Factorization". In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Dis-*

covery and Data Mining. KDD '14. New York, NY, USA: ACM, pp. 115–124. ISBN: 978-1-4503-2956-9.

Ho, Joyce C, Joydeep Ghosh, and Jimeng Sun (2014b). "Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization". In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 115–124.

Ho, Joyce C. et al. (Dec. 1, 2014a). "Limestone: High-throughput candidate phenotype generation via tensor factorization". In: *Journal of Biomedical Informatics*. Special Section: Methods in Clinical Research Informatics 52, pp. 199–211.

Ho, Joyce C et al. (2014b). "Limestone: High-throughput candidate phenotype generation via tensor factorization". In: *Journal of biomedical informatics* 52, pp. 199–211.

Hogan, William R. and Michael M. Wagner (Sept. 1, 1997). "Accuracy of Data in Computer-based Patient Records". In: *Journal of the American Medical Informatics Association* 4.5, pp. 342–355.

Hong S.B.N. et al. (2015). "The 5 phenotypes of high cost patient". In: *Journal of General Internal Medicine* 30 ((Whitman N.; Vakharia N.; Rothberg M.B.) Cleveland Clinic, Cleveland, OH, United States). Number: (Whitman N.; Vakharia N.; Rothberg M.B.) Cleveland Clinic, Cleveland, OH, United States, S73.

Hripcsak, George and David J Albers (2012). "Next-generation phenotyping of electronic health records". In: *Journal of the American Medical Informatics Association* 20.1, pp. 117–121.

Hripcsak, George, David J Albers, and Adler Perotte (July 1, 2015). "Parameterizing time in electronic health record studies". In: *Journal of the American Medical Informatics Association* 22.4, pp. 794–804.

Hripcsak, George et al. (2015). "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers". In: *Studies in health technology and informatics* 216, pp. 574–578.

Hripcsak, George et al. (Aug. 1, 2019). "Facilitating phenotype transfer using a common data model". In: *Journal of Biomedical Informatics* 96, p. 103253.

Hui, Changwei et al. (2015). *Computational Phenotyping via Scalable Bayesian Tensor Factorization*. (Visited on 04/03/2018).

Ipsen, Niels Bruun, Pierre-Alexandre Mattei, and Jes Frellsen (2021). "not-MIWAE: Deep Generative Modelling with Missing not at Random Data". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

J., Cohen (1997). *Design and methods of the Medical Expenditure Panel Survey, household component*. Public Health Service, Agency for Health Care Policy and Research: US Department of Health and Human Services.

Jain, Sachin H. et al. (May 2015). "The digital phenotype". In: *Nature Biotechnology* 33.5. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 5 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group Subject_term: Genetics;Health care;Technology Subject_term_id: genetics;health-care;technology, pp. 462–463.

James, S. A. et al. (1987). "Socioeconomic Status and John Henryism and Hypertension in Blacks and whites." In: *Am J Epidemiol* 126, pp. 664–673.

Johnson, Alistair EW et al. (2016). "MIMIC-III, a freely accessible critical care database". In: *Scientific data* 3, p. 160035.

Joshi, Shalmali et al. (2016). "Identifiable phenotyping using constrained non-negative matrix factorization". In: *arXiv preprint arXiv:1608.00704*.

Ka, Wallston (2005). "The Validity of the Multidimensional Health Locus of Control Scales". In: *J Health Psychol* 10, pp. 623–631.

Kagawa, Rina et al. (July 2017). "Development of Type 2 Diabetes Mellitus Phenotyping Framework Using Expert Knowledge and Machine Learning Approach." In: *Journal of diabetes science and technology* 11.4. Number: 4, pp. 791–799.

Kamkar, Iman et al. (Feb. 1, 2015). "Stable feature selection for clinical prediction: Exploiting ICD tree structure using Tree-Lasso". In: *Journal of Biomedical Informatics* 53, pp. 277–290.

Kho, Abel N et al. (Mar. 1, 2012). "Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study". In: *Journal of the American Medical Informatics Association* 19.2, pp. 212–218.

Kim, M. T. et al. (2021). "Psychosocial phenotyping as a personalization strategy for chronic disease self-management interventions". In: *Am J Transl Res* 13, pp. 1617–1635.

Kim, Yejin et al. (Aug. 2017). "Federated Tensor Factorization for Computational Phenotyping." In: *KDD : proceedings. International Conference on Knowledge Discovery & Data Mining* 2017, pp. 887–895.

King E. et al. (2014). "Early hospital readmissions phenotype: Center level analysis in kidney transplantation". In: *Transplantation* 98 ((King E.; Mcadams-DeMarco M.; Chow E.; Segev D.) Department of Surgery, Johns Hopkins University, Baltimore, MD, United States). Number: (King E.; Mcadams-DeMarco M.; Chow E.; Segev D.) Department of Surgery, Johns Hopkins University, Baltimore, MD, United States, p. 69.

Kingma, Diederik P. and Max Welling (2014). "Auto-Encoding Variational Bayes". In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun.

Kirby, Jacqueline C. et al. (Nov. 2016). "PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability". In: *Journal of the American Medical Informatics Association: JAMIA* 23.6. Number: 6, pp. 1046–1052.

Kirkpatrick, S. I. and V. Tarasuk (2008). "Food Insecurity in Canada". In: *Can J Public Health* 99, pp. 324–327.

Kobyzev, Ivan, Simon J. D. Prince, and Marcus A. Brubaker (2020). "Normalizing Flows: An Introduction and Review of Current Methods". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1. arXiv: 1908.09257.

Koch, P. A., I. R. Contento, and et al. Gray H. L. (2019). "Food, Health, Choices: curriculum and wellness interventions to decrease childhood obesity in fifth-graders". In: *J Nutr Educ Behav* 51, pp. 440–455.

Koola, Jejo D. et al. (Mar. 9, 2018). "Development of an Automated Phenotyping Algorithm for Hepatorenal Syndrome". In: *Journal of Biomedical Informatics*.

Koola J.D. et al. (2017). "Algorithm for identification of patients at high risk for cirrhosis from administrative data". In: *Hepatology* 66 ((Dever J.; Ho S.B.) Gastroenterology, VA San Diego Healthcare System, San Diego, CA, United States). Number: (Dever J.; Ho S.B.) Gastroenterology, VA San Diego Healthcare System, San Diego, CA, United States, 325A.

Kotfila, Christopher and Özlem Uzuner (Dec. 1, 2015). "A systematic comparison of feature space effects on disease classifier performance for phenotype identification of five diseases". In: *Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data* 58, S92–S102.

Kroenke, K. et al. (2010). "The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review". In: *Gen Hosp Psychiatry* 32, pp. 345–359.

Kumar V. et al. (2014). "Natural language processing improves phenotypic accuracy in an electronic medical record cohort of type 2 diabetes and cardiovascular disease". In: *Journal of the American College of Cardiology* 63.12. Number: 12, A1359.

Kummer B.R. et al. (2017). "An electronic health record phenotype of ischemic stroke using non-claims clinical data and machine learning". In: *Stroke* 48 ((Kummer B.R.) Dept of Biomedical Informatics, Columbia Univ Med Cntr, Feil Family Brain and Mind Rsch Institute, New York, NY, United States). Number: (Kummer B.R.) Dept of Biomedical Informatics, Columbia Univ Med Cntr, Feil Family Brain and Mind Rsch Institute, New York, NY, United States.

Lachman, M. E. and Weaver Sl (1998). "Sociodemographic variations in the sense of control by domain: Findings from the MacArthur studies of midlife". In: *Psychol Aging* 13, pp. 553–562.

Landerman, R. et al. (1989). "Alternative models of the stress buffering hypothesis". In: *Am J Community Psychol* 17, pp. 625–642.

Lasko, Thomas A., Joshua C. Denny, and Mia A. Levy (June 24, 2013). "Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data". In: *PLOS ONE* 8.6. Number: 6, e66341.

Lee, Meredith et al. (2020). *WiDS (Women in Data Science) Datathon 2020: ICU Mortality Prediction*. Type: dataset.

Lee, Y. J. et al. (2015). "The association between online health information–seeking behaviors and health behaviors among Hispanics in New York City: a community-based cross-sectional study". In: *J Med Internet Res* 17.

Levoska M. et al. (2017). "Deep phenotyping of patients with xeroderma pigmentosum and trichothiodystrophy". In: *Journal of Investigative Dermatology* 137.5. Number: 5, S46.

Li, Lei, Qin Zhang, and Danfeng Huang (Nov. 2014). "A Review of Imaging Techniques for Plant Phenotyping". In: *Sensors* 14.11. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute, pp. 20078–20111.

Li, Li et al. (Oct. 28, 2015). "Identification of type 2 diabetes subgroups through topological analysis of patient similarity". In: *Science Translational Medicine* 7.311. Number: 311, 311ra174.

Li, Liam et al. (Mar. 15, 2020). *A System for Massively Parallel Hyperparameter Tuning*. arXiv: `1810.05934[cs,stat]`.

Liao, Katherine P. et al. (Aug. 2010). "Electronic medical records for discovery research in rheumatoid arthritis". In: *Arthritis Care & Research* 62.8. Number: 8, pp. 1120–1127.

Liao K.P. et al. (2015). "Development of phenotype algorithms using electronic medical records and incorporating natural language processing". In: *BMJ (Online)* 350 ((Liao

K.P., kliao@partners.org; Karlson E.W.) Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, MA, United States). Number: (Liao K.P., kliao@partners.org; Karlson E.W.) Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, MA, United States.

Lin, Chen et al. (Apr. 1, 2015). "Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record". In: *Journal of the American Medical Informatics Association* 22 (e1), e151–e161.

Lingren, Todd et al. (2016). "Developing an Algorithm to Detect Early Childhood Obesity in Two Tertiary Pediatric Medical Centers". In: *Applied Clinical Informatics* 07.3. Publisher: Schattauer GmbH, pp. 693–706.

Lipton, Zachary C., David Kale, and Randall Wetzel (Dec. 10, 2016). "Directly Modeling Missing Data in Sequences with RNNs: Improved Classification of Clinical Time Series". In: *Machine Learning for Healthcare Conference*. Machine Learning for Healthcare Conference, pp. 253–270.

Lipton, Zachary C., David C. Kale, and Randall C. Wetzel (Oct. 26, 2015). "Phenotyping of Clinical Time Series with LSTM Recurrent Neural Networks". In: *arXiv:1510.07641 [cs]*. arXiv: 1510.07641.

Loi, Michele (Mar. 1, 2019). "The Digital Phenotype: a Philosophical and Ethical Exploration". In: *Philosophy & Technology* 32.1, pp. 155–171.

Lor, M. et al. (2019). "Association Between Health Literacy and Medication Adherence Among Hispanics with Hypertension". In: *J Racial Ethn Health Disparities* 6, pp. 517–524.

Lu, Hsin-Min, Chih-Ping Wei, and Fei-Yuan Hsiao (Apr. 1, 2016). "Modeling healthcare data using multiple-channel latent Dirichlet allocation". In: *Journal of Biomedical Informatics* 60, pp. 210–223.

Mahner, Martin and Michael Kary (May 7, 1997). "What Exactly Are Genomes, Genotypes and Phenotypes? And What About Phenomes?" In: *Journal of Theoretical Biology* 186.1, pp. 55–63.

Marin, G. et al. (1987). "Development of a Short Acculturation Scale for Hispanics". In: *Hisp J Behav Sci* 9, pp. 183–205.

Marlin, Benjamin M. et al. (Jan. 28, 2012). "Unsupervised pattern discovery in electronic health care data using probabilistic clustering models". In: *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. IHI '12. New York, NY, USA: Association for Computing Machinery, pp. 389–398. ISBN: 978-1-4503-0781-9.

Masterson Creber, R. M. et al. (2017). "Identifying the Complexity of Multiple Risk Factors for Obesity Among Urban Latinas". In: *J Immigr Minor Health* 19, pp. 275–284.

Mattei, Pierre-Alexandre and Jes Frellsen (May 24, 2019). "MIWAE: Deep Generative Modelling and Imputation of Incomplete Data Sets". In: *International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 2640-3498. PMLR, pp. 4413–4423.

Matthay, E. C., E. Hagan, and et al. Gottlieb L. M. (2021). "Powering population health research: Considerations for plausible and actionable effect sizes". In: *SSM - Popul Health* 14.10078, p. 9.

Matthews, K. A., L. C. Gallo, and Taylor Se (2010). "Are psychosocial factors mediators of socioeconomic status and health connections: A progress report and blueprint for the future". In: *Ann N Y Acad Sci* 1186, pp. 146–173.

Mayhew, Michael B. et al. (Feb. 1, 2018). "Flexible, cluster-based analysis of the electronic medical record of sepsis with composite mixture models". In: *Journal of Biomedical Informatics* 78, pp. 33–42.

Mayr, Ernst (1973). "The Recent Historiography of Genetics". In: *Journal of the History of Biology* 6.1. Ed. by R. C. Olby and W. B. Provine. Publisher: Springer, pp. 125–154.

McGraw, Deven, Kristen Rosati, and Barbara Evans (2012). "A policy framework for public health uses of electronic health data". In: *Pharmacoepidemiology and Drug Safety* 21 (S1). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pds.2319, pp. 18–22.

Miotto, Riccardo et al. (May 17, 2016a). "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records". In: *Scientific Reports* 6.1. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Outcomes research;Translational research Subject_term_id: outcomes-research;translational-research, p. 26094.

Miotto, Riccardo et al. (2016b). "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records". In: *Scientific reports* 6, p. 26094.

Moore, L. V. et al. (2008). "Associations of the Local Food Environment with Diet Quality—A Comparison of Assessments based on Surveys and Geographic Information Systems: The Multi-Ethnic Study of Atherosclerosis". In: *Am J Epidemiol* 167, pp. 917–924.

Morley, Katherine I. et al. (Nov. 4, 2014). "Defining Disease Phenotypes Using National Linked Electronic Health Records: A Case Study of Atrial Fibrillation". In: *PLOS ONE* 9.11. Publisher: Public Library of Science, e110900.

Mueller, Andreas (2019). *WordCloud*. `https://github.com/amueller/word_cloud`.

Murphy, Shawn N et al. (Mar. 1, 2010). "Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)". In: *Journal of the American Medical Informatics Association* 17.2, pp. 124–130.

Murray, I. and Salakhutdinov Rr (2008). "Evaluating probabilities under high-dimensional latent variable models". In: *Adv Neural Inf Process Syst* 21, pp. 1137–1144.

Nazábal, Alfredo et al. (Nov. 1, 2020). "Handling incomplete heterogeneous data using VAEs". In: *Pattern Recognition* 107, p. 107501.

Neal, Radford M. (Sept. 1993). *Probabilistic Inference Using Markov Chain Monte Carlo Methods*. Technical Report CRG-TR-93-1. Department of Computer Science, University of Toronto, p. 144.

— (2011). "MCMC Using Hamiltonian Dynamics". In: *Handbook of Markov Chain Monte Carlo*. Num Pages: 50. Chapman and Hall/CRC. ISBN: 978-0-429-13850-8.

Nelson, D., G. Kreps, and et al. Hesse B. (2004). "The Health Information National Trends Survey (HINTS): Development, Design, and Dissemination". In: *J Health Commun* 9, pp. 443–460.

Newton, Katherine M et al. (June 1, 2013). "Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network". In: *Journal of the American Medical Informatics Association* 20 (e1), e147–e154.

Ng, Andrew and Michael Jordan (2002). "On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes". In: *Advances in Neural Information Processing Systems*. Vol. 14. MIT Press.

Nicholson, Amanda et al. (Feb. 22, 2013). "Optimising Use of Electronic Health Records to Describe the Presentation of Rheumatoid Arthritis in Primary Care: A Strategy for Developing Code Lists". In: *PLOS ONE* 8.2. Publisher: Public Library of Science, e54878.

Orphanou, Kalia et al. (May 1, 2018). "Incorporating repeating temporal association rules in Naïve Bayes classifiers for coronary heart disease diagnosis". In: *Journal of Biomedical Informatics* 81, pp. 74–82.

Overhage, J Marc et al. (Jan. 1, 2012). "Validation of a common data model for active safety surveillance research". In: *Journal of the American Medical Informatics Association* 19.1, pp. 54–60.

Papamakarios, George, Theo Pavlakou, and Iain Murray (June 14, 2018). "Masked Autoregressive Flow for Density Estimation". In: *arXiv:1705.07057 [cs, stat]*. arXiv: 1705.07057.

Papamakarios, George et al. (Apr. 8, 2021). *Normalizing Flows for Probabilistic Modeling and Inference*. arXiv: 1912.02762[cs,stat].

Pedregosa, F. et al. (2011a). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pedregosa, Fabian et al. (2011b). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85, pp. 2825–2830.

Perros, I. et al. (Nov. 2015). "Sparse Hierarchical Tucker Factorization and Its Application to Healthcare". In: *2015 IEEE International Conference on Data Mining*. 2015 IEEE International Conference on Data Mining, pp. 943–948.

Perros, Ioakeim et al. (2017). "SPARTan: Scalable PARAFAC2 for Large & Sparse Data". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '17. New York, NY, USA: ACM, pp. 375–384. ISBN: 978-1-4503-4887-4.

Petersen, Laura A. et al. (Sept. 1, 1999). "Positive predictive value of the diagnosis of acute myocardial infarction in an administrative database". In: *Journal of General Internal Medicine* 14.9, pp. 555–558.

Pilkonis, P. A. et al. (2011). "Item Banks for Measuring Emotional Distress From the Patient-Reported Outcomes Measurement Information System (PROMIS®): Depression, Anxiety, and Anger". In: *Assessment* 18, pp. 263–283.

Pivovarov, Rimma et al. (Dec. 2015a). "Learning Probabilistic Phenotypes from Heterogeneous EHR Data". In: *Journal of Biomedical Informatics* 58, pp. 156–165.

Pivovarov, Rimma et al. (2015b). "Learning probabilistic phenotypes from heterogeneous EHR data". In: *Journal of biomedical informatics* 58, pp. 156–165.

research, Contento Ir: Nutrition education: linking (2016). *theory, and practice. Third edition.* Burlington, Massachusetts: Jones Bartlett Learning.

Rezende, Danilo Jimenez, Shakir Mohamed, and Daan Wierstra (June 18, 2014). "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". In: *Proceedings of the 31st International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 1938-7228. PMLR, pp. 1278–1286.

Richesson, Rachel L et al. (2013). "Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory". In: *Journal of the American Medical Informatics Association* 20.e2, e226–e231.

Richesson, Rachel L et al. (2016). "Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods". In: *Artificial intelligence in medicine* 71, pp. 57–61.

Ritchie, Marylyn D. et al. (Apr. 9, 2010). "Robust Replication of Genotype-Phenotype Associations across Multiple Diseases in an Electronic Medical Record". In: *The American Journal of Human Genetics* 86.4, pp. 560–572.

Robert, Christian P, George Casella, and George Casella (2010). *Introducing monte carlo methods with r*. Vol. 18. Springer.

Rodriguez, V. A. and A. Perotte (2019). In: *Phenotype inference with semi-supervised mixed membership models*, pp. 304–324.

Roqueiro, Damian et al. (June 15, 2015). "In silico phenotyping via co-training for improved phenotype prediction from genotype". In: *Bioinformatics* 31.12, pp. i303–i310.

Rothman, A. J. and P. Sheeran (2020). *The operating conditions framework: Integrating mechanisms and moderators in health behavior interventions*. Health Psychol.

Rotmensch, Maya et al. (July 20, 2017). "Learning a Health Knowledge Graph from Electronic Medical Records". In: *Scientific Reports* 7.1. Number: 1, p. 5994.

Rubin, Donald B. (1976). "Inference and Missing Data". In: *Biometrika* 63.3. Publisher: [Oxford University Press, Biometrika Trust], pp. 581–592.

Ruffini, Matteo, Ricard Gavalda, and Esther Limon (Nov. 6, 2017). "Clustering Patients with Tensor Decomposition". In: *Machine Learning for Healthcare Conference*. Machine Learning for Healthcare Conference, pp. 126–146.

Russo C. et al. (2017). "Heart failure phenotyping by latent class analysis identifies subpopulations at high risk of mortality and readmissions: Insights from a real world database". In: *Journal of the American College of Cardiology* 69.11. Number: 11, p. 778.

Sa, James (1994). "John Henryism and the health of African-Americans". In: *Cult Med Psychiatry* 18, pp. 163–182.

Saria, Suchi, Andrew Duchi, and Daphne Koller (2011). "Discovering Deformable Motifs in Continuous Time Series Data". In: *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*. IJCAI'11. Barcelona, Catalonia, Spain: AAAI Press, pp. 1465–1471. ISBN: 978-1-57735-514-4.

Saria, Suchi, Daphne Koller, and Anna Penn (2010). "Learning individual and population level traits from clinical temporal data". In.

Savova, Guergana K et al. (Sept. 1, 2010). "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications". In: *Journal of the American Medical Informatics Association* 17.5, pp. 507–513.

Schmiedeskamp, Mia et al. (Nov. 2009). "Use of International Classification of Diseases, Ninth Revision Clinical Modification Codes and Medication Use Data to Identify Nosocomial Clostridium difficile Infection". In: *Infection Control & Hospital Epidemiology* 30.11. Publisher: Cambridge University Press, pp. 1070–1076.

Schulam, Peter, Fredrick Wigley, and Suchi Saria (Feb. 21, 2015). "Clustering Longitudinal Clinical Marker Trajectories from Electronic Health Data: Applications to Phenotyping and Endotype Discovery". In: *Twenty-Ninth AAAI Conference on Artificial Intelligence*. Twenty-Ninth AAAI Conference on Artificial Intelligence.

Scriver, C. R. (2004). "After the genome—the phenome?" In: *Journal of Inherited Metabolic Disease* 27.3, pp. 305–317.

Sepulveda-Pacsi, A. L. and S. Bakken (2017). "Correlates of Dominicans' Identification of Cancer as a Worrisome Health Problem". In: *J Immigr Minor Health* 19, pp. 1227–1234.

Shang, N., C. Liu, and et al. Rasmussen L. V. (2019). "Making work visible for electronic phenotype implementation: Lessons learned from the eMERGE network". In: *J Biomed Inform* 99.10329, p. 3.

Shivade, Chaitanya et al. (Mar. 1, 2014). "A review of approaches to identifying patient phenotype cohorts using electronic health records". In: *Journal of the American Medical Informatics Association* 21.2, pp. 221–230.

Stanhope, K. K., M. Picon, and et al. Schlusser C. (2021). "Chronic Stress and Preconception Health Among Latina Women in Metro Atlanta". In: *Matern Child Health J* 25, pp. 1147–1155.

Suresh, Harini, Peter Szolovits, and Marzyeh Ghassemi (Mar. 20, 2017). "The Use of Autoencoders for Discovering Patient Phenotypes". In: *arXiv:1703.07004 [cs]*. arXiv: 1703.07004.

Swinburn, Boyd A., Gary Sacks, and et al. Kevin D. Hall (2011). "Obesity 1: The global obesity pandemic: shaped by global drivers and local environments". In: *The Lancet* 378, p. 804.

Tamang, Suzanne and Simon Parsons (2011). "Using Semi-parametric Clustering Applied to Electronic Health Record Time Series Data". In: *Proceedings of the 2011 Workshop on Data Mining for Medicine and Healthcare*. DMMH '11. New York, NY, USA: ACM, pp. 72–75. ISBN: 978-1-4503-0843-4.

Teixeira P.L. et al. (2017). "Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals". In: *Journal of the American Medical Informatics Association* 24.1. Number: 1, pp. 162–171.

Thorpe, Lorna E et al. (June 15, 2006). "Study Design and Participation Rates of the New York City Health and Nutrition Examination Survey, 2004". In: *Preventing Chronic Disease* 3.3, A94.

Tran, Truyen et al. (Apr. 2015). "Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM)." In: *Journal of biomedical informatics* 54, pp. 96–105.

Turner, Clayton A. et al. (Aug. 22, 2017). "Word2Vec inversion and traditional text classifiers for phenotyping lupus." In: *BMC medical informatics and decision making* 17.1. Number: 1, p. 126.

Urteaga, I., M. McKillop, and N. Elhadad (2020). "Learning endometriosis phenotypes from patient-generated data". In: *NPJ Digit Med* 3, pp. 1–14.

Vazquez Guillamet, Rodrigo et al. (Nov. 17, 2016). "Chronic obstructive pulmonary disease phenotypes using cluster analysis of electronic medical records." In: *Health informatics journal*.

Wagholikar, Kavishwar B et al. (May 15, 2020). "Polar labeling: silver standard algorithm for training disease classifiers". In: *Bioinformatics* 36.10, pp. 3200–3206.

Wallach, H. M. et al. (2009). In: *Evaluation methods for topic models*, pp. 1105–1112.

Wang, Yichen et al. (2015a). "Rubik: Knowledge Guided Tensor Factorization and Completion for Health Data Analytics". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15. New York, NY, USA: ACM, pp. 1265–1274. ISBN: 978-1-4503-3664-2.

Wang, Yichen et al. (2015b). "Rubik: Knowledge guided tensor factorization and completion for health data analytics". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 1265–1274.

Ware, J. E. et al. (2001). "How to score and interpret single-item health status measures: a manual for users of the SF-8 health survey". In: *Linc RI Qual Inc* 15, p. 5.

Ware, John E. and Cathy Donald Sherbourne (1992). "The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual Framework and Item Selection". In: *Medical Care* 30.6, pp. 473–483.

Watson, James D. (1970). *Molecular biology of the gene*. Second edition. OCLC: ocm00098773. New York: W. A. Benjamin. 662 pp. ISBN: 978-0-8053-9603-4.

Wei, Wei-Qi et al. (Nov. 13, 2010). "A high throughput semantic concept frequency based approach for patient identification: a case study using type 2 diabetes mellitus clinical

notes." In: *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2010, pp. 857–861.

Weiskopf, Nicole G. et al. (Oct. 1, 2013). "Defining and measuring completeness of electronic health records for secondary use". In: *Journal of Biomedical Informatics* 46.5, pp. 830–836.

Weiskopf, Nicole Gray and Chunhua Weng (2013). "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research". In: *Journal of the American Medical Informatics Association : JAMIA* 20.1. Number: 1, pp. 144–151.

Weiss, B. D., M. Z. Mays, and et al. Martz W. (2005). "Quick assessment of literacy in primary care: the newest vital sign". In: *Ann Fam Med* 3, pp. 514–522.

Xiao, Y. et al. (2020). "epigenetics, and adipose tissue metabolism in the obese state". In: *Nutr Metab* 17, p. 88.

Xu, Haowen et al. (May 31, 2019). "On the Necessity and Effectiveness of Learning the Prior of Variational Auto-Encoder". In: *arXiv:1905.13452 [cs, stat]*. arXiv: 1905.13452.

Xu, Jie et al. (Nov. 1, 2015). "Review and evaluation of electronic health records-driven phenotype algorithm authoring tools for clinical and translational research". In: *Journal of the American Medical Informatics Association* 22.6, pp. 1251–1260.

Xu, Zhenxing et al. (Feb. 1, 2020). "Identifying sub-phenotypes of acute kidney injury using structured and unstructured electronic health record data with memory networks". In: *Journal of Biomedical Informatics* 102, p. 103361.

Yj, Lee (2013). *Online Health Information Seeking Behaviors of Hispanics in New York City*. Columbia University.

Yoon, S., N. Suero-Tejeda, and S. Bakken (2015). "A data mining approach for examining predictors of physical activity among urban older adults". In: *J Gerontol Nurs* 41, pp. 14–20.

Yoon, S., A. B. Wilcox, and S. Bakken (2013). "Comparisons among Health Behavior Surveys: Implications for the Design of Informatics Infrastructures That Support Comparative Effectiveness Research". In: *EGEMs* 1.

Yu, Sheng et al. (Apr. 1, 2017). "Surrogate-assisted feature extraction for high-throughput phenotyping". In: *Journal of the American Medical Informatics Association* 24 (e1), e143–e149.

Yu, Sheng et al. (Jan. 1, 2018). "Enabling phenotypic big data with PheNorm". In: *Journal of the American Medical Informatics Association* 25.1, pp. 54–60.

Zhang, Yichi et al. (Dec. 2019). "High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP)". In: *Nature Protocols* 14.12. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 12 Primary_atype: Protocols Publisher: Nature Publishing Group Subject_term: Bioinformatics;Biomarkers;Diseases;Software Subject_term_id: bioinformatics;biomarkers;diseases;software, pp. 3426–3444.

Zhao, Juan et al. (Oct. 2019). "Detecting time-evolving phenotypic topics via tensor factorization on electronic health records: Cardiovascular disease case study". In: *Journal of Biomedical Informatics* 98, p. 103270.

Zheng, Tao et al. (Jan. 2017). "A machine learning-based framework to identify type 2 diabetes through electronic health records." In: *International journal of medical informatics* 97, pp. 120–127.

Zhou, Jiayu et al. (2014). "From Micro to Macro: Data Driven Phenotyping by Densification of Longitudinal Electronic Medical Records". In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '14. New York, NY, USA: ACM, pp. 135–144. ISBN: 978-1-4503-2956-9.

Zhou, Shang-Ming et al. (2016). "Defining Disease Phenotypes in Primary Care Electronic Health Records by a Machine Learning Approach: A Case Study in Identifying Rheumatoid Arthritis." In: *PloS one* 11.5. Number: 5, e0154515.

Zhou, Zhi-Hua (Jan. 1, 2018). “A brief introduction to weakly supervised learning”. In: *National Science Review* 5.1, pp. 44–53.

Zou, H. and Hastie T. (2005). “Regularization and variable selection via the elastic net”. In: *J R Stat Soc Ser B Stat Methodol* 67, pp. 301–320.