Exploiting Structured Data For Learning Contagious Diseases Under Incomplete Testing

Maggie Makar [1] **Lauren West** [2] **David Hooper** [2] **Eric Horvitz** [3] **Erica Shenoy** [2] **John Guttag** [1]

Abstract

One of the ways that machine learning algorithms can help control the spread of an infectious disease is by building models that predict who is likely to become infected making them good candidates for preemptive interventions. In this work we ask: can we build reliable infection prediction models when the observed data is collected under limited, and biased testing that prioritizes testing symptomatic individuals? Our analysis suggests that when the infection is highly transmissible, incomplete testing might be sufficient to achieve good out-of-sample prediction error. Guided by this insight, we develop an algorithm that predicts infections, and show that it outperforms baselines on simulated data. We apply our model to data from a large hospital to predict *Clostridioides difficile* infections; a communicable disease that is characterized by both symptomatically infected and asymptomatic (i.e., untested) carriers. Using a proxy instead of the unobserved untested-infected state, we show that our model outperforms benchmarks in predicting infections.

1. Introduction

Preemptively identifying individuals at a high risk of contracting a contagious (i.e., transmissible) infection is important for guiding treatment decisions to mitigate symptoms, and for preventing further spread of the infection through appropriate isolation. In this paper, we study how to build individual-level predictive models for contagious infections while explicitly addressing the challenges inherent to contagious diseases.

Building accurate infection prediction models is hindered by two main factors. First, contagious infections defy the usual *iid* assumption central to most machine learning methods. This is because an individual's infection state is not independent of their contacts' infection states. Previous work has often relied on expert knowledge to construct exposure proxies (Wiens et al., 2012; Oh et al., 2018). It is then assumed that conditional on the exposure proxy and individual characteristics, individual outcomes are independent of one another. Such an assumption is violated if the exposure proxy is noisy or misspecified leading to inaccurate prediction.

Second, the observed data is biased. The primary clinical purpose of testing for a disease is to provide guidance for treatment decisions for the individual being tested. Therefore, there is a strong bias in who is tested—people for whom knowing whether they have the disease will affect treatment (e.g., symptomatic individuals) are far more likely to be tested than other members of the population. But for some infectious diseases, only a fraction of those individuals carrying the pathogen experience noticeable symptoms. We use the term "incomplete testing" to describe the scenario where only a small, biased subset of individuals harboring a pathogen are tested. Incomplete testing makes learning accurate models difficult since the collected labels are missing not at random, leading to biased and inconsistent estimates.

In this work, we leverage the non-independence of outcomes to construct robust predictors. Specifically, we use the knowledge that infections are caused by exposure to the pathogen through contacts to impute missing infection labels. Our proposed approach uses the fact that an individual's infection state provides useful information about their contacts' true infection states. This information is used to generate pseudo-labels for untested individuals, mitigating issues caused by incomplete testing. The key idea behind our approach is that highly structured patterns of disease transmission can serve as a complementary signal to identify even untested carriers. The stronger that signal is, the less impact that incomplete testing will have. Our contributions can be summarized as follows:

1. We identify two properties of collected data that can be exploited to mitigate the effects of incomplete testing.
2. We present an algorithm that leverages that insight to

[1]CSAIL, MIT [2]Infection Control Unit, Massachusetts General Hospital [3]Microsoft. Correspondence to: Maggie Makar <mmakar@mit.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

predict the probability of an untested individual carrying the disease.

3. We empirically evaluate the effectiveness of our method on both simulated data and real data for a common and morbid contagious disease. We show that predictions from our model can be used to inform efficient testing and isolation policies. Using EHR data from a large hospital, we show that our model outperforms baselines on the task of predicting a healthcare associated infection.

2. Related work

Infectious disease modeling. Modeling the transmission of infectious diseases has been extensively studied in the epidemiology literature using SIS/SIR models and several other variants (Kermack & McKendrick, 1927). These epidemiological models focus on the *aggregate* levels of infections in a community. This is distinct from our approach, which focuses on predicting individual level infections. In the machine learning literature, previous work has has often relied on expert knowledge to construct exposure proxies (Wiens et al., 2012; Oh et al., 2018). It is then assumed that conditional on the exposure proxy and individual characteristics, individual outcomes are independent of one another. Similar to our approach, (Fan et al., 2016) and (Makar et al., 2018) take into account structured data, namely contact networks to compute infection estimates (Fan et al., 2016; Makar et al., 2018). We differ from these approaches in that (1) we do not make parametric assumptions about the joint distribution of the observed or latent variables, and instead use nonparametric models (neural networks) to model the infection states, (2) we do not assume all infections will become symptomatic as is done in (Fan et al., 2016), and (3) unlike the approach taken by (Makar et al., 2018), we model time evolving sequences of infections taking into account the exposure states of potential asymptomatic carriers.

Semi-supervised learning. Our proposed approach relies on transductive reasoning to generate labels for untested individuals. In that, it is closely related to semi-supervised learning methods, such as pseudo-labeling (Lee, 2003), and self-training (Robinson et al., 2020). In traditional pseudo-labeling, the transductive power comes from the fact that points similar to each other in the input space have similar outputs. Here, the rich structure in the data allows for more: we can construct pseudo-labels for untested individuals not just by relying on their similarity to other labeled instances, but also by observing their observed contacts' infection states. Our empirical results, and analysis are similar in spirit to concepts presented in the semi-supervised literature, specifically the cluster assumption (Seeger, 2000; Rigollet, 2007), which we discuss later.

Graph Neural Networks. Our proposed approach incorporates knowledge of the contact network. In that it is similar to Graph Neural Networks (GNNs), which utilize relational data to generate prediction estimates (Zhou et al., 2018). GNNs fall into two categories. The first relies on transductive reasoning and cannot generalize to new communities (e.g., (Kipf & Welling, 2017)). The second relies on inductive reasoning, which can be used to generate estimates for previously unseen graphs (e.g., (Hamilton et al., 2017)). Our work is similar to the latter category, with an important distinction: our approach leverages unlabeled data giving more accurate, and robust estimates.

Our work can be viewed as combining the strengths of semi-supervised learning, and GNNs to address limited testing. Our approach augments the strengths of those two approaches with ideas from domain shift and causal inference, such as importance weighting (Cortes et al., 2010) to address biased testing.

3. Problem setting

Setup. Let $y^t \in \{0, 1\}$ denote an individual's true infection/carrier state at time t, with $y^t = 1$ if an individual is symptomatically infected or asymptomatically carrying the pathogen, and 0 otherwise. For brevity, we will refer to y^t as the true infection state at time t. We use $\overline{\mathbf{x}}^t \in \mathcal{X}^t$ to denote a vector of the individual's features at time t, and define J_i^t to be the set of indices of i's contacts at time t. We assume that the contact network is fully observed, i.e., that the contact indices are known. We note that the assumption of fully observed networks is less likely to be violated in the context of hospital associated infections, where the majority of patients' interactions and contacts are routinely recorded, compared to community acquired infections. Our results on real data show that even with incomplete networks, our approach outperforms others.

Let $e_i^t \in \mathbb{R}_{\geq 0}$ denote i's exposure state at time t, with $e_i^t = \sum_{j \in J_i^t} y_j^t$. The exposure state is fully observed only when all of i's contacts have been tested, but otherwise either partially observed or unobserved. Define $\mathbf{x}^t = \overline{\mathbf{x}}^t || e^t$, where $||$ is the concatenation operator, i.e., $\mathbf{x}^t \in \mathcal{X}^t \times \mathbb{R}_{\geq 0}$. Let $o^t \in \{0, 1\}$ denote the observation state, with $o^t = 1$ if an individual's label is observed, i.e., if the individual has been tested for the infection. We use the super-script $:t$ to denote variables from time $t = 0$ up to and including time t, e.g., $\mathbf{x}^{:t} = [\mathbf{x}^0, ..., \mathbf{x}^s, ..., \mathbf{x}^t]$.

Throughout, we use capital letters to denote variables, and small letters to denote their values. We use $P(\boldsymbol{X}^t, O^t, Y^{t+1})$ to denote the unknown distribution over the full joint. Under biased testing, we have that $P(\boldsymbol{X}^t | O^t = 1) \neq P(\boldsymbol{X}^t | O^t = 0) \neq P(\boldsymbol{X}^t)$. We assume that $0 < P(O^t = o | \boldsymbol{X}^t = \mathbf{x}) < 1$, for all $\mathbf{x} \in \mathcal{X}$,

and $o \in \{0, 1\}$. This is the same as the overlap assumption in the causality literature. In addition, we assume that i's outcome is conditionally independent of i's contacts given $\mathbf{x}_i$ (which is itself a function of the contacts' outcomes). We consider the case where we have access to (1) a labeled (i.e., tested) set of individuals $\mathcal{D}_1 = \{\mathcal{D}_1^t\}_{t=0}^T = \{(\mathbf{x}_i^t, y_i^t), \dots (\mathbf{x}_{n_1^t}^t, y_{n_1^t}^t)\} \sim P(\boldsymbol{X}^t, Y^{t+1} | O^t = 1)$, and (2) an unlabeled (untested) set of individuals $\mathcal{D}_0 = \{\mathcal{D}_0^t\}_{t=0}^T = \{\mathbf{x}_i^t, \dots, \mathbf{x}_{n_0^t}^t\} \sim P(\boldsymbol{X}^t | O^t = 0)$, such that for each $i \in \mathcal{D}_0 \cup \mathcal{D}_1$, and each $t \in [0, T]$, we have that $J_i^t \in \mathcal{D}_0 \cup \mathcal{D}_1$. We use $\mathcal{U}^t$ to denote the set of indices of untested individuals at time t.

Notation	Meaning
y_i^t	i's infection/carrier state at time t (true infection state for short)
$\overline{\mathbf{x}}_i^t$	i's features at time t
e_i^t	i's (partially) observed exposure state at time t
$\mathbf{x}_i^t$	The concatenation of $\overline{\mathbf{x}}_i^t$, and e_i^t
$\mathbf{x}_i^{:t}$	The collection of an individual's features, and exposure states from time $t = 0$ until $t = t$, i.e., $\mathbf{x}_i^{:t} - [\mathbf{x}_i^0, ..., \mathbf{x}_i^s, ..., \mathbf{x}_i^t]$
J_i^t	The set of indices of i's contacts at time t
o_i^t	Observation state for the infection label. $o_i^t = 1$ if i's infection state is observed at time t (i.e., if i was tested for the infection at time t), and 0 otherwise
$\mathcal{D}_1$	Data $(\mathbf{x}, y$ tuples) for tested individuals
$\mathcal{D}_0$	Data $(\mathbf{x})$ for untested individuals
$w^t(\mathbf{x}_i^t)$	Probability that an individual with characteristics $\mathbf{x}_i^t$ gets tested
$\mathcal{U}^t$	the set of indices of untested individuals at time t.
$\mathcal{A}_i^t$	The set of ancestors of i at time t whose outcomes are unobserved i.e., $\mathcal{A}_i^t = J^t(i) \cap \mathcal{U}^t$

Table 1. Summary of notation

Learning objective. We want to learn $f : \mathbf{x}^{:T} \to y^{T+1}$. To focus the discussion on the novel component of our approach, we first consider a setting in which we predict the outcomes for a single time step: making predictions for $t = 2$, using data from $t = 0, 1$. We drop the time superscript when it can be inferred from the context. We present the full model predicting infection sequences over time in section 5. Let ℓ be the logistic loss. Our goal is to find $f \in \mathcal{F}$, where $\mathcal{F}$ is some hypothesis space such that the risk of incorrectly classifying the infection state $R(f) = \mathbb{E}_{\boldsymbol{X},Y}[\ell(f(\boldsymbol{X}^t), Y^{t+1})]$ is minimized. We first consider a scenario where we have oracle access to the infection states of the untested population, but we return to the more realistic, non-oracle scenario later. Note that having access to the untested population's infection states implies that exposure states are also fully observed (by definition of the exposure states). Under the conditional independence assumption, we can view the risk as a sum of independent losses. Define the inverse probability of being tested as $w^t(\boldsymbol{X}) = {}^{P(O^t = o)}/_{P(O^t = o | \boldsymbol{X}^t)}$, following Robins (1998), and Robins et al. (2000). Because of the overlap assumption, under biased testing we have that:

$$R(f) = R^{w^t}(f) = \mathbb{E}[w^t(\boldsymbol{X})\ell(f(\boldsymbol{X}), Y)]. \tag{1}$$

$R^{w^t}(f)$ cannot be directly computed since the expectation is defined with respect to the unobserved distribution. However, the following reweighted empirical loss is an unbiased estimator of $R^{w^t}(f)$:

$$\varepsilon(f) = \sum_{i \in \mathcal{D}_0^t \cup \mathcal{D}_1^t} w_i^t \ell(f(\mathbf{x}_i^t), y_i^{t+1}),$$

where $w_i^t = {}^{p(O^t = o_i^t)}/_{g(o_i^t | \mathbf{x}_i^t)}$, $p(O^t = o_i^t)$ is the empirical estimate of $P(O^t = o)$, and $g(o_i^t | \mathbf{x}_i^t)$ is the estimated probability of getting tested conditioned on individual characteristics. Without oracle access to untested individuals' infection states, we cannot directly minimize $\varepsilon(f)$ for $i \in \mathcal{D}_0^t$ since their labels are never observed. In addition, without access to untested individuals' infection states, the samples $\mathbf{x}^t \sim P(\boldsymbol{X}^t | O^t = 1)$ are incomplete. This is because $\mathbf{x}_i^t$ includes e_i^t, which is a function of $y_j^t : j \in J_i^t$. We only fully observe e_i^t, and hence $\mathbf{x}_i^t$ for individuals whose contacts have all been tested. To address this, we define Q as the set of all possible distributions over y_i^t for $i \in \mathcal{D}_0^t$. Our risk is now defined with respect to both Q, and f.

Let $\hat{y}_i \sim Q$, $\hat{e}_i^t = \sum_{j \in J^t(i)} \mathbb{1}\{j : o_j^t = 1\} \cdot y_j^t + \mathbb{1}\{j : o_j^t = 0\} \cdot \hat{y}_j^t$, $\hat{\mathbf{x}}_i = \overline{\mathbf{x}}_i^t || \hat{e}_i^t$, and $\hat{w}_i^t = {}^{p(O = o_i)}/_{g(\hat{\mathbf{x}}_i, o_i)}$, our task is to find Q and f, such that the following empirical risk is minimized:

$$\begin{aligned} \varepsilon(f, Q) = &\sum_{i \in \mathcal{D}_1^t} \hat{w}_i^t \ell(f(\hat{\mathbf{x}}_i^t), y_i^{t+1}) \\ &+ \sum_{i \in \mathcal{D}_0^t} \hat{w}_i^t \ell(f(\hat{\mathbf{x}}_i^t), \hat{y}_i^{t+1}). \end{aligned} \tag{2}$$

We next consider how to leverage properties of the problem to efficiently minimize $\varepsilon(f, Q)$.

4. Exploiting structure as a regularizer

We seek to constrain the candidate sets $\mathcal{F}$ and Q to avoid overfitting. To do so, we exploit both the interdependence among individuals' infection states and the availability of unlabeled data. Recall that the exposure state of an individual is the sum of that individual's contacts' infection states. This means that when we draw $\hat{y}_i^t$ from Q, *we are implicitly drawing the exposure states for i's contacts', by definition*

of $\hat{e}_i^t$. This becomes obvious if we decompose $\hat{e}_i^t$ as follows: $\hat{e}_i^t = \sum_{j \in J^t(i)} \mathbb{1}\{j : o_j^t = 1\} \cdot y_j^t + \mathbb{1}\{j : o_j^t = 0\} \cdot \hat{y}_j^t$, and $\hat{y}_j^t \sim Q$. This decomposition immediately implies two properties that should hold for "good" Q's. First, Q should assign infection states, $\hat{y}_i^t$, that are consistent with i's contacts' infection states. Consider the case where two individuals i, and j came into contact with each other at time t. Suppose that j tests positive at time $t+1$. For simplicity, suppose that i, and j have no contacts other than each other. Here Q should assign i a high probability of infection because in order to become infected j must have been exposed to the pathogen through i. Second, note that Q is assigning pseudo-labels for the infection states of untested contacts, this means that Q's imputed labels should be similar to the labels predicted by f. A good regularization method should then explicitly encourage the pseudo-labels to be similar to the estimated labels from f. This intuition is encoded in the main loss in our proposed approach:

$$f^*, Q^* = \min_{f,Q} \frac{1}{n_1^t} \sum_{i:o_i^t=1} \hat{w}_i^t \ell(f(\hat{\mathbf{x}}_i^t), y_i^{t+1}) + \frac{\lambda}{|J_i^t \cap \mathcal{U}^t|} \sum_{j \in J_i^t \cap \mathcal{U}^t} \hat{w}_j^{t-1} \ell(\mathbb{1}\{f(\mathbf{x}_j^{t-1}) > \tau\}, \hat{y}_j^t) \tag{3}$$

where $|.|$ denotes the set cardinality, $\lambda \geq 0$, and τ are parameters to be picked using cross validation, $\hat{y}_j^t \sim Q^*$, and $\hat{e}_i^t = \sum_{j \in J^t(i)} \mathbb{1}\{j : o_j^t = 1\} \cdot y_j^t + \mathbb{1}\{j : o_j^t = 0\} \cdot \hat{y}_j^t$. When $\lambda > 0$, this objective is somewhat similar to pseudo-labeling (Lee, 2003), it would encourage the imputed infection states, $\hat{y}_j^t \sim Q$, to conform with the prediction from f. When $\lambda = 0$, equation 3 prioritizes finding good predictions for the labeled samples, ignoring possible structure implied by the data. Note that in the second term in equation 3, we have $f(\mathbf{x}_j^{t-1})$, rather than $f(\hat{\mathbf{x}}_j^{t-1})$, meaning we assume no imputed exposure component for contacts at time $t-1$. This is because we are considering the simple setting where $t-1=0$, i.e., $t-1$ is the beginning of the observation period and no exposure has happened yet. We later consider more complicated settings where the contacts' inputs also include an exposure state.

4.1. When does structure work as a regularizer?

We now ask: when do we expect equation 3 to yield models superior to those that ignore structure? First, if the untested-infected individuals' contacts are more likely to become infected compared to the untested-uninfected individuals' contacts. We stress that we do not need all of the untested-uninfected individuals' contacts to be infected and all of the untested-uninfected contacted to be uninfected. We only require that there is *separation* between the likely outcomes of two groups' contacts. Intuitively, in such a setting, the contacts' outcomes provide a signal revealing the untested individuals' true infection states. In practice, high separability should occur, even in settings of low and biased testing, assuming the observed data satisfies a property we refer to as the **potency property**. The potency property can be viewed as an extension of the margin condition in classification (Tsybakov et al., 2004; Audibert et al., 2007). It implies that infections cluster so that infected-untested individuals tend to have more infected contacts than do uninfected-untested individuals. Such a condition will be satisfied if the infection is sufficiently transmissible.

Second, even if there is high separability but $\hat{\mathbf{x}}$ makes it difficult to identify a learnable mapping from $\hat{\mathbf{x}}$ to the imputed $\hat{y}$, minimizing equation 3 instead of the objective on only the labeled data does not help. Such is the case when untested-healthy and untested-infected individuals "look" the same, meaning they have very similar characteristics and exposure states. This property is often referred to as the cluster assumption in semi-supervised learning literature (Rigollet, 2007; Seeger, 2000). The cluster assumption states that individual characteristics, and exposure states tend to form near discrete clusters, with homogeneous labels within each cluster. Intuitively, it means that we can learn the correct clustering of individuals that separates infected from uninfected individuals, up to a permutation of the labeling. We refer to this property as the **dissimilarity property**.

The degree to which these two properties are satisfied in the observed data will depend largely upon the infection being studied and the environment in which it is spreading. However, as we show in section 6, even when these properties do not hold, our proposed approach performs as well as the best baseline. I.e., even in the worst case scenario, the regularization "does no harm."

5. Proposed method

Our proposed model, a Model for Infections under Incomplete Testing (MIINT) leverages labeled and unlabeled data to predict infections over time. MIINT minimizes a variant of equation 3, which is modified to predict the spread of infection over an arbitrary time horizon. Let $\mathcal{A}_i^t$, be the set of ancestors of i at time t whose outcomes are unobserved, i.e., $\mathcal{A}_i^t = J^t(i) \cap \mathcal{U}^t$, $\mathcal{A}^{t-1} = \bigcup_{j \in \mathcal{A}_i^t} J^{t-1}(j) \cap \mathcal{U}^{t-1}$, etc. The loss at time t is defined as:

$$\mathcal{L}^t = \frac{1}{n_1^t} \sum_{i \in \mathcal{D}_1} \hat{w}_i^t \ell(f(\hat{\mathbf{x}}_i^t), y_i^{t+1}) + \sum_{s=0}^{t} \frac{\lambda}{|\mathcal{A}_i^t|} \sum_{j \in \mathcal{A}_i^s} \hat{w}_j^s \ell(\mathbb{1}\{f(\hat{\mathbf{x}}_j^s) > \tau\}, \hat{y}_j^s), \tag{4}$$

and the objective is to find f^*, Q^*, such that:

$$f^*, Q^* = \min_{f,Q} \frac{1}{T} \sum_t \mathcal{L}^t.$$

It is possible to consider the family of candidate functions $\mathcal{F}$ to be any family of non-parametric estimators. For our implementation, we take $\mathcal{F}$ to be the space of recurrent neural networks (RNNs). We assume that f does not vary over time (though that is an assumption that could be relaxed). We propagate the predicted state forward in time, meaning in practice f takes in $\mathbf{x}^t, e^t$ and $\hat{y}^t$ to predict $\hat{y}^{t+1}$. This ensures that exposures at time $< t$ are taken into account when predicting infections at time t. Note that equation 4 can be decomposed into the independent sums of individual losses, as well as their ancestors' losses. This means we can use stochastic gradient descent, with gradient updates defined with respect to mini-batches, as is typically done. One limitation is that equation 4 as stated would require keeping track of all the ancestors' states since $t = 0$, which can be prohibitive for long observation periods. In practice, one would consider a subset of $\mathcal{A}_i^t$ based upon the properties of the disease being studied.

The algorithm used to train MIINT, similar to pseudo-labeling (Lee, 2003), is an expectation maximization algorithm, where we iterate between computing the expected label for the untested samples (i.e., finding the optimal $\hat{Q}$), and finding the optimal f that maximize the likelihood of the observed labels under $\hat{Q}$ until convergence. Convergence is achieved when the change in loss defined over the samples with observed labels in a held out validation set is $< \epsilon$ for some small ϵ. For our purposes, we find it sufficient to let Q be a deterministic function rather than an actual distribution. However, our approach is extendable to allow Q to be a distribution, for example using techniques described in (Tran et al., 2017).

Finally, recall that we need to estimate $\hat{w}_i^t = {}^{p(O=o_i)}/_{g(\hat{\mathbf{x}}_i, o_i)}$. We follow (Chernozhukov et al., 2017) in using an independent sample to estimate g. Importantly, g depends on $\hat{\mathbf{x}}$. So we follow an iterative process: after every epoch of training, we use the most updated f to estimate the unobserved labels in the independent weighting set. This in turn gives us an estimate for $\hat{e}$ and $\hat{\mathbf{x}}$ for the independent weighting sample. We use these imputed values to learn an updated g. The updated g provides estimates for the weights of the training samples of the main prediction model, which are used to reweight the loss function for the next epoch, and so forth.

6. Experiments

We evaluate our model on a simulated and a real data setting. All models presented in this paper are implemented using Tensorflow (Abadi et al., 2016). Our code is available at `github.com/mymakar/miint`.

In the simulated setting, unlike the real data setting, we have access to the true infection state, which allows us to evaluate the performance of the model and baselines under different patterns of infection. In both settings, we present results from our model (MIINT) and five baselines:

1. Optimistic Model (**OM**): a model that assumes that all unobserved labels are equal to 0,
2. No Exposure Model (**NEM**): a model that ignores exposure, and attempts to predict infections solely based on the individual characteristics,
3. GraphSAGE (**GNN**): a graph neural network that takes into account the contact network, and observed infection states (Hamilton et al., 2017) but ignores untested individuals,
4. Pseudo-Labeling (**PL**): a semi-supervised learning method that takes into account untested individuals but ignores the graph structure (Lee, 2003),
5. ORacle Model (**ORM**): an unattainable model that has oracle access to the true labels for the whole population.

For all baselines, we weight the loss from each individual by the inverse of their estimated propensity to be tested, w_i^t, which is estimated using an independent sample following Chernozhukov et al. (2017). For our model, we use the iterative weighting technique outlined in section 5. For all baselines as well as our approach (MIINT), we keep the neural network architecture fixed. We use cross-validatation to get the values of λ, and τ. Results from unweighted models and details about cross-validation and network architecture are included in the supplement.

6.1. Simulation experiments

The simulation experiments demonstrate how MIINT can be used to inform testing and isolation policies that lead to reduction in infection rates, as well as empirically validate our conjectures regarding the conditions under which MIINT is expected to perform better than other methods.

Setup. We simulate a world in which there are three types of people: symptomatic and infected if exposed (G_0), asymptomatic (but carrier) if exposed (G_1), and non-infected/non-carriers (G_2). If exposed, individuals in group G_0 become infected and symptomatic, hence they are more likely to get tested. If exposed, individuals in group G_1 become infected without displaying symptoms. This group is unlikely to get tested. Finally, individuals in G_2 are unlikely to get the infection or carry the pathogen even if exposed. To simulate individuals' characteristics (i.e., $\overline{\mathbf{x}}$), we map the distinct groups to distinct MNIST digits. We use MNIST images because they can be easily classified as visually similar or dissimilar, which enables us to design experiments where

the dissimilarity property can be manipulated, as described later.

Let ν_i denote the pixels of an MNIST image i. For G_0 we randomly sample without replacement $n/3 \cdot T$ elements from the set $\{\nu_i\}_{i:d_i=0}$, where n is the total sample size. We do the same for G_1, and G_2 but here we sample from $\{\nu_i\}_{i:d_i=1}$, and $\{\nu_i\}_{i:d_i=2}$, respectively. Note that the infection states will be different within each group, since infection also depends on the exposure state. We draw the edge sets $\{J^t(i)\}_{i\in n, t\in[0,T]}$ according to a stochastic block model, parameterized by the matrix B, where $B_{k,l}$ is the probability that an individual from G_k forms an edge with an individual from G_l. B is important in simulating different levels of carrier potency. When $B_{1,k}/B_{1,2}$ for $k = \{0, 1\}$ approaches 1, members of the asymptomatic carrier group are equally likely to form an edge with individuals who are susceptible to symptomatic infections (G_0) as with individuals who are non-infected/non-carriers (G_2). This is a low-separation setting, which is unfavorable for our approach. On the other hand, if $B_{1,k}/B_{1,2} = 5$, for example, individuals in G_0, and G_1 are 5 times more likely to form an edge with someone in a susceptible group as compared to forming an edge with an individual in G_2. This a favorable high-separation setting.

We mimic the situation where testing started after a significant proportion of the population has been exposed by randomly setting the true exposure state of 20% of the population to be 1 at time $t = 0$. Exposure for each individual $e_i^t = \sum_{j\in J_i^t} y_j^t \geq 1$. The true infection label $y_i^{t+1} = \mathbb{1}\{i \in (G_0, G_1)\} \cdot \mathbb{1}\{e_i^t = 1\}$. We introduce noise by randomly flipping the labels of 1% of the population. If an individual tests positive at $t < T$, their label remains positive until $t = T$. We define $p_{\texttt{obs}}$ to be the proportion tested (their true label is observed). We pick the probability of observing i's label based on i's true infection state, meaning, $p(o_i|y_i = 1) \neq p(o_i|y_i = 0)$. For all the simulations, we set $T = 6$ and we draw 500×6 samples for each of the training, validation, and testing sets. We simulate an independent sample to compute the weights w_i, so we also draw 500×6 samples that are used to train and validate weighting models. For each experiment, we draw 10 different datasets, and report the mean and standard deviation of the performance metric across the 10 draws.

Informing testing and isolation policies. Here, we highlight how our model can inform efficient testing and isolation policies. We simulate biased and limited testing by setting $p(o_i|y_i = 1)/p(o_i|y_i = 0) = 5$, and $p_{\texttt{obs}} = .1$ respectively. We set $B_{1,k}/B_{1,2} = 5$, making it a high potency setting where MIINT is expected to perform well. We mimic a situation where no isolation interventions are taken at training time. At deployment time, we fix a testing budget of at most $p_{\texttt{test}}\%$ of the total population on each time step.

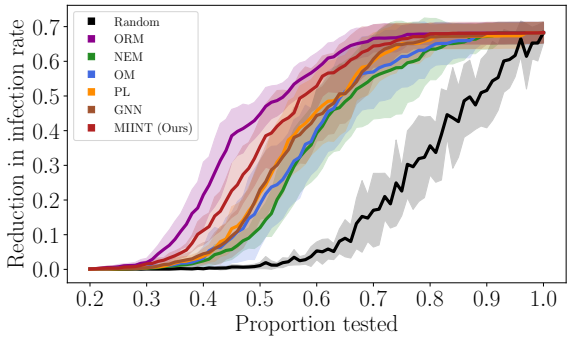

Figure 1. Reduction in infection rates relative to a policy that does not isolate infections (no-action policy) as the daily testing budget varies. Our model achieves the highest reductions in policy relative to all realistic (i.e., non-oracle) models.

We use the predictions from each model to inform who gets tested by picking the top $p_{\texttt{test}}\%$ with the highest predicted probability of infection. Of those tested, individuals who are infected are "isolated" by setting their edges for the subsequent time steps to 0. They are also taken out of the population eligible for further testing.

We compute the infection rate based on the isolation policy suggested by each model at the end of the time horizon, i.e., at time T. For a model M, the infection rate $\pi_M = n^{-1} \cdot \sum_i y_i^T$. We define π_0 as the infection rate under a no-action policy, that is if no isolation interventions are taken. Our main metric of interest is the reduction in infection rate relative to the no-action policy $= (\pi_0 - \pi_M)/\pi_0$. Figure 1 shows the reduction in infection rate on the $y-$axis for different values of the testing budget $p_{\texttt{test}}\%$ on the $x-$axis. In addition to the main baselines, we also show results from a random testing policy. The results show that for each testing budget, our model outperforms all feasible baselines leading to uncovering more individuals who should be isolated, thus achieving a higher reduction in infection rates. The results imply that our model is able to achieve near oracle infection control with 70% testing, compared to $\approx 90\%$ for the baselines.

In the next two settings, we empirically validate our conjectures about the two properties that enable our model to outperform others, and explore what happens as these favorable properties are weakened to the point of non-existence.

Sensitivity to the potency property. Here, we fix $p_{\texttt{obs}} = .1$ and $p(o_i|y_i = 1)/p(o_i|y_i = 0) = 5$, and sweep over carrier potency by varying the value of $B_{1,k}/B_{1,2}$ from 1 (low potency) to 5 (high potency). Figure 2(top) shows $B_{1,k}/B_{1,2}$ on the $x-$axis and the AUROC on the $y-$axis. The plot shows that MIINT outperforms other baselines when there is high potency, and as potency declines, its performance becomes similar to that of the other baselines. This supports our conjecture that our regularization approach is advantageous

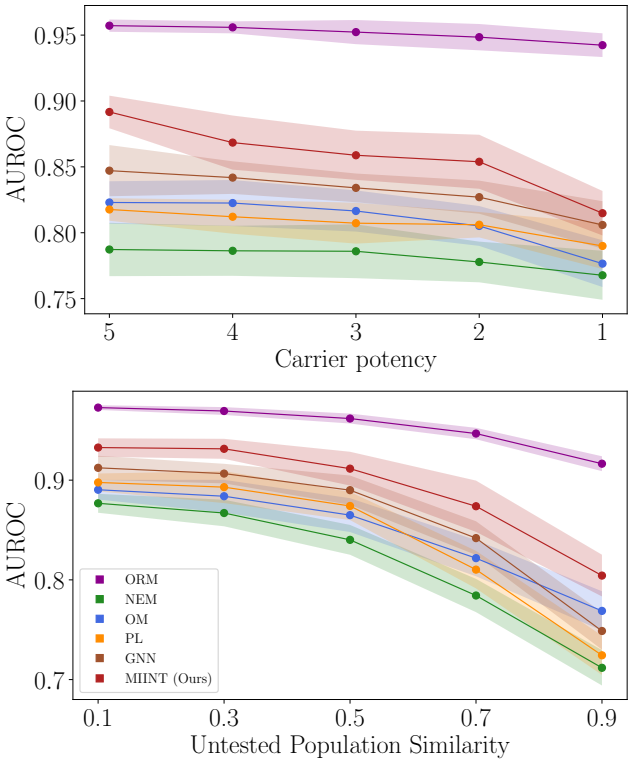

Figure 2. Top: Impact of varying levels of carrier potency controlled by $^{B_{1,k}}/_{B_{1,2}}$. Our model outperforms baselines, especially in cases with high potency. Bottom: Impact of high (=.9) and low (=.1) similarity between the characteristics of the untested-uninfected and untested-infected populations. Our model outperforms baselines when the two populations are dissimilar.

when the true infection states for an individual is strongly related to their contacts' infection states.

Sensitivity to the dissimilarity property. Here we examine what happens when the untested-infected individuals have characteristics that are similar to untested-uninfected individuals. We do so by moving the untested, and possibly infected[1] individuals to "look" similar to the non-infected/non-carrier individuals. Specifically, we sample pairs of images $\{(\nu_i, \nu_j)\}_{i,j:d_i=1,d_j=2}$. We then use VoxelMorph (Balakrishnan et al., 2018), a learning-based framework for deformable, pairwise image registration to learn a function that gives us a deformation field which we then apply it to pairs of images, moving ν_i to look more similar to ν_j. Using VoxelMorph in this way allows us to control the degree of similarity between images.

Figure 3 shows a sample image morphing for a pair of images using VoxelMorph.

Figure 2(bottom) shows the results of this setting. The $x-$axis can be viewed as the degree of similarity between the two untested groups with 0 being dissimilar (i.e., the original images without any deformation) and 1 being very similar (i.e., all images of the digit 1 look almost identical to 2's). The $y-$axis is the average AUROC. We see that all models perform worse as members in G_1 look more and more similar to those in G_2. We also see that MIINT outperforms all baselines when the two groups are dissimilar, and performs as well as the others when the mapping from input space to label becomes more difficult.

The last two experiments confirm our conjectures about the properties necessary for MIINT to perform well, and imply that MIINT "does no harm": at worst it performs comparably to alternatives, and at best it can give significantly better performance. Additional results examining the effect of bias and limited testing are in the supplement.

6.2. Real data experiment

Here, our task is to predict the onset of *Clostridioides difficile*, (*C. difficile*) infections among patients in a large urban hospital. *C. difficile* infections are contagious bacterial infections that attack the gut, and cause over 300,000 infections annually in the US (Magill et al., 2014). As with most contagious infections, asymptomatic carriers of *C. difficile* exist and can contribute to the spread of the infection (Riggs et al., 2007).

Setup. Using EHRs, we extract daily characteristics of patients who were admitted to the hospital between 09/01/2012 and 06/01/2014. We follow similar inclusion criteria as (Oh et al., 2018; Makar et al., 2018), outlined in detail in the supplement. We collect all patient characteristics available upon admission (e.g., gender, age, medical history) as well as daily characteristics (e.g., lab tests). We collect contact networks, where an edge exists if two patients are in the same room on the same day or if they came into contact with the same nurse on the same day.

Here, we have partial access to the true infection states, since not all the patients are tested, making accurate evaluation of different models impossible. Therefore, we exploit the hospital's testing protocols to construct a proxy "true" label and a proxy "observed" label. Whether a patient is diagnosed as *C. difficile* positive or not is a result of two, or possibly three tests. First, an enzyme immunoassay (EIA) and Glutamate dehydrogenase (GDH) test are conducted. If the results of the two tests are discordant, a polymerase chain reaction (PCR) assay acts as a tie-breaker. Previous studies comparing the outcomes of the two groups (those who have concordant, positive EIA/GDH results vs. discordant EIA/GDH and PCR positive) have shown that the former experiences more severe complications (Origüen et al., 2018; Polage et al., 2015). This suggests that concordant, positive EIA/GDH can act a proxy for symptomatic

[1]Individuals in G_1 are only infected if they get exposed.

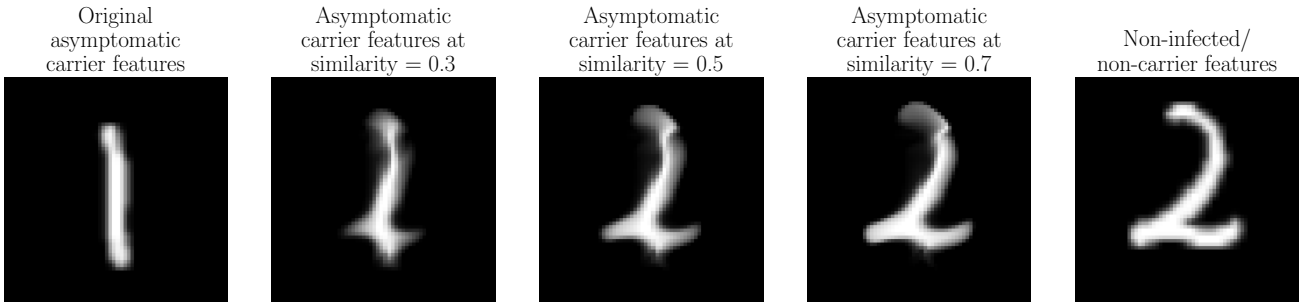

Figure 3. Varying similarity between asymptomatic carrier features and non-infected/non-carrier features using VoxelMorph (Balakrishnan et al., 2018)

infections whereas discordant EIA/GDH, and PCR positive may be identifying patients who are carriers but have low toxin levels and are therefore asymptomatic.

In this experiment, we hide the EIA/GDH discordant lab tests at training time, presenting them as untested individuals to all models. At the final evaluation time, however, we compare the models' predictions to the true infection labels. The true infection state is determined to be positive for patients who tested positive through a concordant, positive EIA/GDH or a discordant EIA/GDH test followed by a positive PCR test. True negatives are defined similarly; they are patients who tested negative either through a concordant, negative EIA/GDH, or a discordant EIA/GDH and a negative PCR test. In addition to the baselines outlined in section 6.1, we allow one of the models full access to the EIA/GDH discordant, PCR positive/negative labels, and refer to it as a "partial oracle" model (POM) since it has access to the PCR positive/negative labels, but not the full infection states. The latter are unavailable because the vast majority of patients in the hospital are not tested. We also compare our results to a state-of-the-art prediction model for *C. difficile* infections (Wiens et al., 2012), which is a logistic regression model that takes into account the varying importance of different risk factors over the hospitalization, and relies on medical knowledge to construct exposure proxies. We refer to this model as the Expert driven Logistic Regression (ELR).

We split the data into 5 subsets based on time. The first subset holds 6 months of data and is used to train the main infection prediction models. The second and third subsets contain 5 months of data each, and are used for validation and testing of the main prediction model. The last 2 subsets are used for training and validation of the weighting models, and each contain 2 months worth of data. We report the AUROC, the True Positive Rate (TPR) at the threshold which achieves a False Positive Rate (FPR) of 10% on the test set.

Table 2 shows the results of the models on the test set. For several models, the unweighted model outperforms its weighted counterpart. We show the better performing version here, and index it with "–U" to denote that it is the unweighted version. Results from all models, and results broken down by concordant EIA/GDH as well as discordant EIA/GDH are in the supplement. Standard deviations are calculated by taking 100 bootstrap replicates of the test set data. We see that MIINT outperforms almost all others on both reported metrics. The one exception is ELR: MIINT and ELR achieve comparable AUROCs but MIINT has a significantly better TPR. MIINT outperforms POM even though the latter has access to better labels. We hypothesize that this is because in addition to accurately predicting the tested patients, MIINT is also capturing truly untested infections, and utilizing these estimates to accurately impute the exposures of the concordant EIA/GDH patients as well as the discordant EIA/GDH, PCR positive patients.

	TPR@ FPR=10%	AUROC
POM	0.49 (0.014)	0.73 (0.003)
NEM–U	0.45 (0.009)	0.7 (0.006)
OM–U	0.45 (0.012)	0.7 (0.005)
ELR	0.53 (0.008)	**0.82 (0.006)**
GNN	0.24 (0.005)	0.59 (0.005)
PL–U	0.58 (0.012)	0.78 (0.006)
MIINT–U	**0.6 (0.007)**	0.81 (0.006)

Table 2. Performance metrics for *C. difficile* infection prediction on the test set.

7. Conclusion

We presented MIINT, a model that predicts contagious infections. Unlike other models, MIINT works well even when labels are generated using biased and limited testing. It does so by exploiting the fact that, in practice, data related to contagious diseases are not *i.i.d*. The key idea is that structured patterns of infection transmission can serve as a complementary signal to identify even untested carriers. The stronger that signal is, the less impact that biased and

incomplete testing will have.

We identified two properties that determine the extent to which MIINT outperforms other approaches. The first states that the more transmissible the infection, the better MIINT performs. The second is the degree to which characteristics of untested and infected individuals and characteristics of the untested and healthy individuals form discrete clusters– an important property in general for semi-supervised learning.

We showed empirically that MIINT can be used to inform testing and isolation strategies that can reduce total infections. We also showed that even if the two properties outlined above are absent, MIINT still performs well. In an experiment using EHR data, we showed that MIINT outperforms baselines when when used to predict CDI.

Future work and limitations There are several future directions that extend the work presented here. One direction is extensions to hypergraphs. The "flat" graphs used in this paper allow each individual to be connected to others through one edge only. Hypergraph extensions would allow each individual to be connected to others through multiple edges, each encoding a different mode of contact (e.g., contact through clinician sharing, or room sharing). Encoding multiple relationships through hypergraphs could enable the identification of different transmission routes.

Because of the obvious relevance of our work to the current pandemic we should note that our model is best-suited for infections which spread through close contact where contact tracing is available through structured EHR data. Extensions to our approach that address community-acquired infections must take into account incomplete and imperfect contact tracing.

In conclusion, we believe this work is a first step down an important path. If predictive models are to play a useful role in limiting the spread of contagious infections, they must take into account the interdependence of outcomes, and the fact that untested individuals are capable of spreading the disease before they have been diagnosed.

Acknowledgements

We are thankful for feedback from the members of the Clinical and Applied Machine Learning group at MIT. We thank the members of the MGH data team for curating the medical date. We are also thankful for the anonymous reviewers who gave insightful and thorough comments. This work was funded by Quanta Computer and Microsoft.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: A system for large-scale machine learning. In *Symposium on Operating Systems Design and Implementation*, pp. 265–283, 2016.

Audibert, J.-Y., Tsybakov, A. B., et al. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2): 608–633, 2007.

Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., and Dalca, A. V. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9252–9260, 2018.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017.

Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pp. 442–450, 2010.

Fan, K., Li, C., and Heller, K. A unifying variational inference framework for hierarchical graph-coupled hmm with an application to influenza infection. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 3828–3834, 2016.

Hamilton, W., Ying, Z., and Leskovec, J. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pp. 1024–1034, 2017.

Kermack, W. O. and McKendrick, A. G. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Lee, D.-H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. 2003.

Magill, S. S., Edwards, J. R., Bamberg, W., Beldavs, Z. G., Dumyati, G., Kainer, M. A., Lynfield, R., Maloney, M., McAllister-Hollod, L., Nadle, J., et al. Multistate point-prevalence survey of health care–associated infections.

New England Journal of Medicine, 370(13):1198–1208, 2014.

Makar, M., Guttag, J., and Wiens, J. Learning the probability of activation in the presence of latent spreaders. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Oh, J., Makar, M., Fusco, C., McCaffrey, R., Rao, K., Ryan, E. E., Washer, L., West, L. R., Young, V. B., Guttag, J., et al. A generalizable, data-driven approach to predict daily risk of clostridium difficile infection at two large academic health centers. *infection control & hospital epidemiology*, 39(4):425–433, 2018.

Origüen, J., Corbella, L., Orellana, M., Fernandez-Ruiz, M., Lopez-Medrano, F., San Juan, R., Lizasoain, M., Ruiz-Merlo, T., Morales-Cartagena, A., Maestro, G., et al. Comparison of the clinical course of clostridium difficile infection in glutamate dehydrogenase-positive toxin-negative patients diagnosed by pcr to those with a positive toxin test. *Clinical Microbiology and Infection*, 24(4): 414–421, 2018.

Polage, C. R., Gyorke, C. E., Kennedy, M. A., Leslie, J. L., Chin, D. L., Wang, S., Nguyen, H. H., Huang, B., Tang, Y.-W., Lee, L. W., et al. Overdiagnosis of clostridium difficile infection in the molecular test era. *JAMA internal medicine*, 175(11):1792–1801, 2015.

Riggs, M. M., Sethi, A. K., Zabarsky, T. F., Eckstein, E. C., Jump, R. L., and Donskey, C. J. Asymptomatic carriers are a potential source for transmission of epidemic and nonepidemic clostridium difficile strains among long-term care facility residents. *Clinical infectious diseases*, 45(8):992–998, 2007.

Rigollet, P. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8(Jul):1369–1392, 2007.

Robins, J. 1997 proceedings of the section on bayesian statistical science. 1998.

Robins, J. M., Hernan, M. A., and Brumback, B. Marginal structural models and causal inference in epidemiology, 2000.

Robinson, J., Jegelka, S., and Sra, S. Strength from weakness: Fast learning using weak supervision. *arXiv preprint arXiv:2002.08483*, 2020.

Seeger, M. Learning with labeled and unlabeled data. Technical report, 2000.

Tran, D., Ranganath, R., and Blei, D. Hierarchical implicit models and likelihood-free variational inference. In *Advances in Neural Information Processing Systems*, pp. 5523–5533, 2017.

Tsybakov, A. B. et al. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1): 135–166, 2004.

Wiens, J., Horvitz, E., and Guttag, J. V. Patient risk stratification for hospital-associated c. diff as a time-series classification task. In *Advances in Neural Information Processing Systems*, pp. 467–475, 2012.

Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*, 2018.