Exploiting Structured Data For Learning Contagious Diseases Under Incomplete Testing: Appendix

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

A. Architecture and cross-validation

The core network architecture is kept constant for all models and experiments outlined in the paper. All models have a core recurrent neural network (RNN). The RNN takes in the individual characteristics, the infection state from the previous time point, and an estimate of the exposure (except for the No exposure model, NEM, which ignores exposure). For our model (MIINT), the exposure estimate is based on the imputed values according to Q, whereas for the optimistic model (OM) it is the sum of observed neighbor infection states from the previous time point (assuming untested is uninfected). For the Oracle model (ORM), the estimate of exposure is the sum of true neighbor infection states from the previous time point.

The RNN inputs are passed through one fully dense layer with a $\tanh$ activation, giving an intermediate layer of dimension = 64 units. The output is then passed to another 64-unit dense layer which is finally passed through a sigmoid to give the final probability of infection.

For the simulation experiment, we found that different values of τ did not affect the prediction much. So we set $\tau = .5$, but pick λ based on cross-validation using a grid of values = Cross-validation is done to pick the value of λ from the candidate values $[0, 1e^{-1}, 1e^{-2}, 1, 1e^{1}, 1e^{2}, 1e^{3}]$. This is done via 2-fold cross validation. We pick the final value to be the one that maximizes the AUROC defined with respect to the observed labels in a held out validation set.

Cross-validation for the real data experiment is similar, though her we found that values of τ are important, so in addition to λ, we also pick the value of τ from the candidate values $[0.001, 0.01, 0.1, 0.5]$.

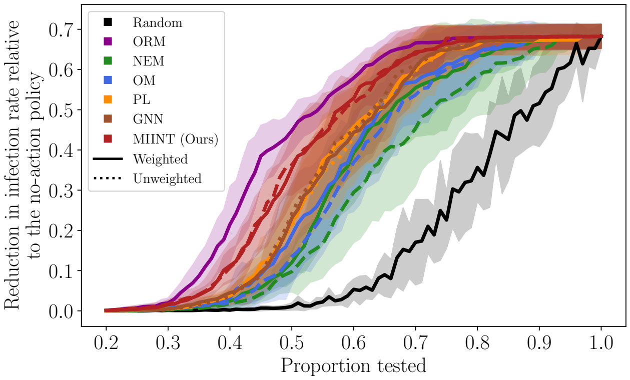

Figure 4. Reduction in infection rates relative to a policy that does not isolate infections (no-action policy) as the daily testing budget varies. Our model achieves the highest reductions in policy relative to all realistic (i.e., non-oracle) models.

B. Additional simulation results

B.1. Unweighted model results

Figures 4, 5, and 6 show the same results as figure 1, 2(left), and 2(right) respectively but here all plots include the results from unweighted variants of the models presented in the main text.

B.2. Additional simulation settings

Impact of biased testing. To explore the impact of biased testing under favorable conditions, we create high potency by setting $B_{1,k}/B_{1,2} = 5$ for $k = \{0, 1\}$. We set $p_{\text{obs}} = .1$, and sweep over the odds of testing conditional on group membership. Results are shown in figure 7, where the $x-$axis shows the odds of testing $(= p(o_i|y_i = 1)/p(o_i|y_i = 0))$, and the $y-$axis shows the AUROC on the held-out test set, averaged over 10 simulations. We see that the weighted version of MIINT outperforms all others. This happens because NEM completely ignores exposure, OM assumes that 90% of the population $(1 - p_{\text{obs}})$ has $y_i = 0$ (which affects its estimate of e^t), whereas MIINT tries to impute the labels for those 90% based on their neighbors infection states. Here the difference between OM and NEM is not large because $p_{\text{obs}} = .1$, which is very low.

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

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

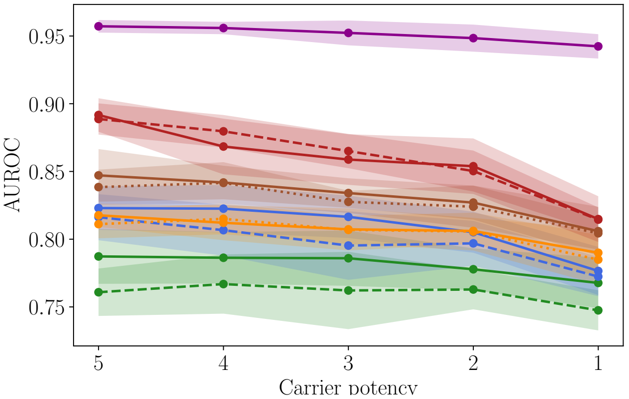

Figure 5. Impact of varying levels of carrier potency controlled by $B_{1,k}/B_{1,2}$. Our model outperforms baselines, especially in cases with high potency.

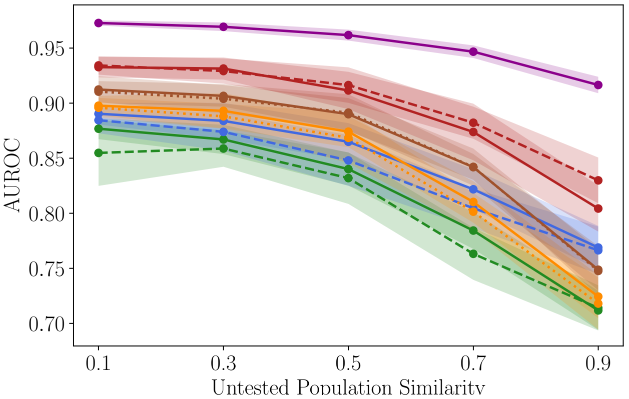

Figure 6. Impact of high (=.9) and low (=.1) similarity between the characteristics of the untested-healthy and untested-infected populations. Our model outperforms baselines when the two populations are dissimilar.

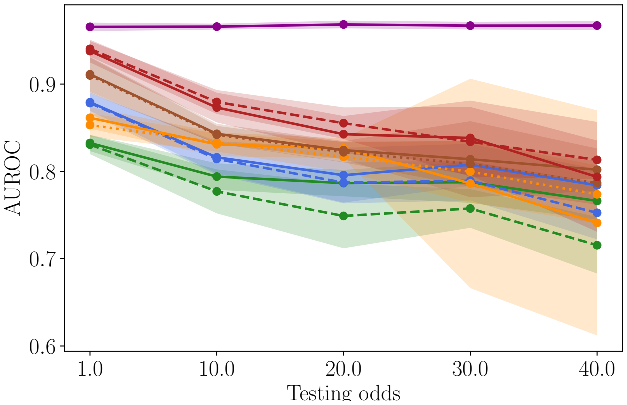

Figure 7. Impact of biased testing, $x-$axis shows the odds ratio of testing given characteristics $(= p(o_i|y_i = 1)/p(o_i|y_i = 0))$, 1 implies randomized testing. Our model does better than baselines for most levels of bias, and similar to baselines at extreme bias.

This means that the exposure estimate that OM relies on is a poor estimate. Results from the subsequent experiment highlight that.

In addition, we see that PL has very high variance for highly biased testing. This makes sense because intuitively, PL assigns labels for the untested population by considering similar patients in the tested population. Under highly biased testing, the labeled and unlabeled population are drastically different, making it difficult to generalize to the unlabeled population without leveraging the rich structured data.

Impact of limited testing. The setup for this experiment is similar to the previous one but here we fix the testing odds, $p(o_i|y_i = 1)/p(o_i|y_i = 0) = 5$, and sweep over the level of testing $p_{\texttt{obs}}$. Figure 8 shows the results, with $p_{\texttt{obs}}$ on the $x-$axis and the AUROC on the $y-$axis, averaged over 10 simulations. We see that weighted MIINT performs as well as the other models at the two extremes of testing levels, and does better at all other levels of testing. Here we see that OM outperforms NEW when the level of testing is sufficiently high, which is expected since OM inherently assumes no unobserved infections. As the testing levels increase, that assumption becomes more correct. The performance of NEM also improves with higher levels of testing since it has access to a cleaner y label, however, it is never able to perform as well as MIINT or OM because it does not take exposure as an input.

C. Real data

Inclusion Criteria. Similar to (Oh et al., 2018; Makar et al., 2018), we exclude all hospitalizations of patients younger than 18. We do so because predicting pediatric *C. difficile* infection is a significantly different task from that of the adult population. We also exclude patients with

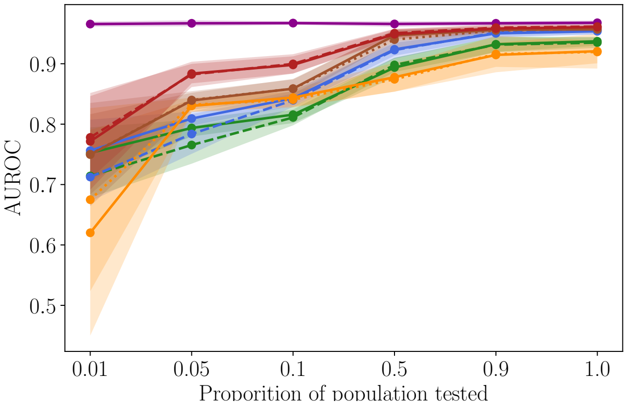

Figure 8. Impact of limited testing. Our model does better than baselines at every level of testing. Our model achieves near oracle accuracy at low levels of testing bias, and high proportion tested.

suspected community acquired infections, since predicting nosocomial infections (i.e., hospital associated infections) is a significantly different task that that of community-acquired infections. Again, we follow (Oh et al., 2018; Makar et al., 2018) in defining community acquired infections as those who get the *C. difficile* infection diagnosis in the first 2 days of their visit, and those who have had a *C. difficile* infection infection in the 14 days prior to their hospitalization.

Patient Features. Similar to (Oh et al., 2018; Makar et al., 2018), we include patient demographics, which are available upon admission such as age, gender, number and length of previous hospitalizations, reason and source of visit (e.g., transferred from a Skilled Nursing Facility or admitted through the emergency room). We capture medical history by including all ICD-9 procedure and diagnosis codes from prior visits that happened at most 90 days prior to the main (index) visit. We collect data from the index visit up to one day before the prediction date. This includes medications, lab tests ordered and their results.

D. Additional real data results

The results from all weighted and unweighted models are shown in table 3. Figure 9 shows the precision-recall curve. The figure highlights that our model achieves higher precision at high recall values.

In addition, we also show the same performance metrics defined with respect to the concordant EIA/GDH label only (table 4) and the discordant EIA/GDH label, PCR labels (table 5).

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.

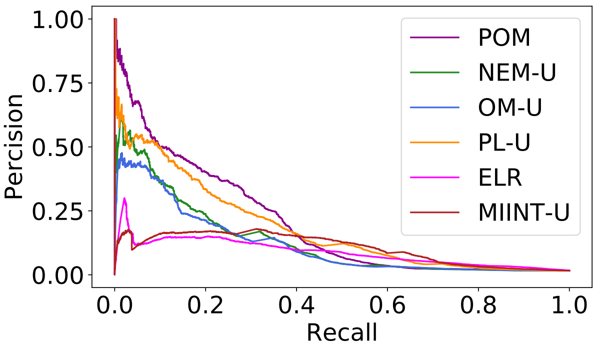

Figure 9. Our model achieves the highest precision at high recall values.

	TPR@ FPR=10%	AUROC
POM	0.49 (0.014)	0.73 (0.003)
NEM	0.33 (0.008)	0.69 (0.006)
NEM–U	0.45 (0.009)	0.7 (0.006)
OM	0.44 (0.008)	0.74 (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)
GNN–U	0.22 (0.007)	0.55 (0.007)
PL	0.26 (0.008)	0.62 (0.005)
PL–U	0.58 (0.012)	0.78 (0.006)
MIINT–U	0.6 (0.007)	0.81 (0.006)
MIINT	0.51 (0.007)	0.78 (0.007)

Table 3. Performance metrics for *C. difficile* infection prediction on the test set, where the target label $=1$ if an individual has a non-discordant GDH/EIA test, or if they have a PCR. positive test.

	TPR@ FPR=10%	AUROC
POM	0.48 (0.014)	0.72 (0.005)
NEM	0.39 (0.018)	0.69 (0.01)
NEM-U	0.54 (0.018)	0.69 (0.009)
OM	0.48 (0.018)	0.72 (0.006)
OM-U	0.52 (0.015)	0.69 (0.004)
ELR	0.6 (0.02)	0.85 (0.006)
GNN	0.34 (0.016)	0.62 (0.011)
GNN–U	0.35 (0.021)	0.59 (0.007)
PL	0.37 (0.017)	0.64 (0.011)
PL–U	0.57 (0.01)	0.75 (0.009)
MIINT–U	0.63 (0.016)	0.82 (0.011)
MIINT	0.58 (0.015)	0.79 (0.011)

Table 4. Performance metrics for *C. difficile* infection prediction on the test set, where the target label $=1$ if an individual has positive, concordant GDH/EIA test, and 0 otherwise.

	TPR@	AUROC
	FPR=10%	
POM	0.49 (0.02)	0.73 (0.009)
NEM	0.28 (0.01)	0.68 (0.007)
NEM–U	0.38 (0.01)	0.7 (0.007)
OM	0.4 (0.018)	0.75 (0.007)
OM–U	0.38 (0.019)	0.7 (0.007)
ELR	0.46 (0.014)	0.8 (0.007)
GNN	0.15 (0.011)	0.57 (0.007)
GNN–U	0.12 (0.01)	0.52 (0.003)
PL	0.17 (0.013)	0.61 (0.006)
PL–U	0.57 (0.015)	0.81 (0.006)
MIINT–U	0.57 (0.009)	0.8 (0.007)
MIINT	0.44 (0.013)	0.76 (0.005)

Table 5. Performance metrics for *C. difficile* infection prediction on the test set, where the target label = 1 if an individual has a discordant GDH/EIA test followed by a positive PCR test, and 0 otherwise.

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.