A Data-driven Approach to Identifying Asymptomatic C. diff Cases

University of Iowa compepi computational epidemiology research

Hankyu Jang[1], Philip Polgreen[2], Alberto Segre[1], Daniel K. Sewell[3], Sriram Pemmaraju[1],

[1]Department of Computer Science, [2]Carver College of Medicine, [3]Department of Biostatistics, University of Iowa

** for the CDC MInD-healthcare group*

The University of Iowa

Abstract

Asymptomatic carriers of an infection make it more challenging to understand the characteristics of that infection (e.g., parameters such as R_0) and to design, implement, and evaluate interventions. Asymptomatic carriers are usually not tested, which also means we do not have "ground truth" labels for these cases in our data. In this paper, we propose a 2-stage classification model for inferring asymptomatic carriers of Clostridioides difficile (C. diff) infections (CDI), a common healthcare-associated infection that causes almost half a million illnesses in the US each year. Guided by hypotheses derived from literature on risk factors for C. diff carriers, we design a Stage 1 model for detecting asymptomatic C. diff carriers that is trained on symptomatic CDI cases. We evaluate the performance of this Stage 1 model by designing a Stage 2 model to predict CDI incidence that uses among its inputs exposure to asymptomatic C. diff carriers inferred by our Stage 1 model. Results from this evaluation lead to two findings. First, our results show that the best performing Stage 1 model depends on all of the standard risk factors for CDI except for high-risk antibiotics. This is an intriguing finding that highlights an important difference between the risk profile of CDI patients and C. diff carriers. Second, we show that adding exposure to asymptomatic cases as an input to the Stage 2 CDI classification model leads to better performance. This result implies that asymptomatic C. diff carriers do in fact contribute to CDI spread, confirming an important conjecture from the CDI literature.

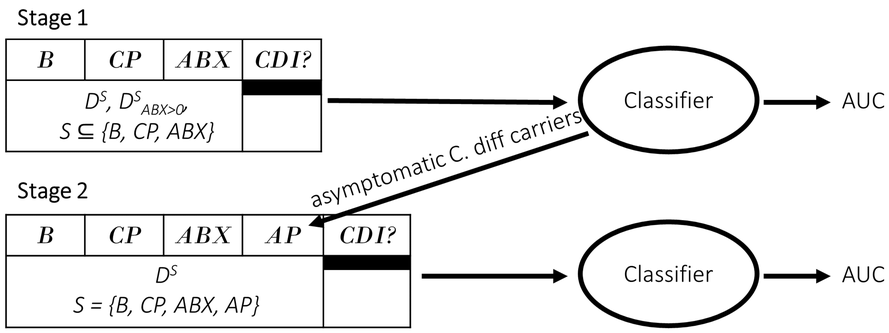
UIHC dataset

The data used in this paper consist of anonymized electronic medical records (EMR) and admission-discharge-transfer records (ADT) for patient visits at the UIHC for the period 2007–2011. The 154,230 patient visits in the data are divided into two groups: (i) $visit_{CDI}$, visits during which patients tested positive for CDI and (ii) $visit_{CDIx}$, the rest of the visits. We exclude short visits where patients are discharged within 48 hours of admission to account for hospital acquired infection. For each visit in $visit_{CDIx}$, we generate one instance per day (CDIx instances) from the admission date to discharge date for that visit. Similarly, we generate daily instances (CDI instances) for each visit in $visit_{CDI}$, starting from the admission date, but only until three days before the CDI positive test date.

Risk factors for CDI

B: Baseline. (i) The length of stay of the visit until the date of the instance, (ii) age, (iii) gender, (iv) previous UIHC visit within 60 days, and the number of gastric acid suppressors prescribed during the visit that are (v) H2-receptor antagonists and (vi) proton pump inhibitors.
ABX: Antibiotics. The number of high-risk antibiotics prescribed during the visit that are (i) Amoxillin or Ampicillin, (ii) Clindamycin, (iii) Third generation Cephalosporin, (iv) Fourth genenration Cephalosporin, and (v) Fluoroquinolone.
CP: *CDI pressure*. For each visit in $visit_{CDI}$ and $visit_{CDIx}$, we keep track of the number of infectious CDI patients in the same room or unit, daily.
AP: *Asymptomatic C. diff carrier pressures*. For each visit in $visit_{CDI}$ and $visit_{CDIx}$, we keep track of the number of asymptomatic C. diff carriers in the same room or unit, daily. These carriers are identified in the Stage 1 Model.

Two stage model



CDI prediction results on Stage 1 and Stage 2 models

The three Stage 2 models that uses *AP* that is computed in Stage 1 model using dataset $D^{B,CP}$ (AUC: 0.733, 0.729, 0.727) outperform the best performing Stage 1 model ($D^{B,ABX,CP}$, AUC: 0.719), clearly indicating that exposure to asymptomatic C. diff carriers impacts the spread of CDI.

Table 2: AUC on Stage 1 models

	D^B	$D^{B,ABX}$	$D^{B,CP}$	$D^{B,ABX,CP}$	$D^B_{ABX>0}$	$D^{B,ABX}_{ABX>0}$	$D^{B,CP}_{ABX>0}$	$D^{B,ABX,CP}_{ABX>0}$
AUC	0.676	0.635	0.704	*0.719	0.594	0.584	0.672	0.648

[a] AUC with asterisk denote best performer for $D^S, S \subseteq \{B, CP, ABX\}$

Table 3: AUC on Stage 2 models

	$D^{B,ABX,CP,AP}$							
AP	D^B	$D^{B,ABX}$	$D^{B,CP}$	$D^{B,ABX,CP}$	$D^B_{ABX>0}$	$D^{B,ABX}_{ABX>0}$	$D^{B,CP}_{ABX>0}$	$D^{B,ABX,CP}_{ABX>0}$
10%	0.712	0.687	*0.733	0.710	0.700	0.724	0.697	0.703
5%	0.701	0.690	*0.727	0.685	0.693	0.714	0.689	0.702
3%	0.689	0.698	*0.729	0.690	0.710	0.704	0.686	0.711

[a] AUC with asterisk denote best performer

Spatio-temporal clustering

G_{CDI}*: CDI case proximity graph.* Nodes are CDI cases at the UIHC during the period 2007–2011 and edges connects pairs of CDI cases in spatio-temporal proximity (cases occurred within 14 days, 30 m apart).
$G_{RCDI\ x\%}$: *revealed CDI case proximity graph*. Nodes are CDI cases and asymptomatic CDI cases output by the Stage 1 model when it was required to mark *x*% of visits in $visit_{CDIx}$ as asymptomatic C. diff carrier visits. We compared the G_{CDI} and $G_{RCDI\ x\%}$ on the four different measures of density. All the density measures were larger in $G_{RCDI\ x\%}$ than in G_{CDI}. The cases in $G_{RCDI\ x\%}$ exhibit spatio-temporal clustering based on our results on statistical tests (e.g. Knox test, we observed the p-value of 0).

Table 6: Network statistics

	G_{CDI}	$G_{RCDI3\%}$	$G_{RCDI5\%}$	$G_{RCDI10\%}$
$\|V\|$	783	4241	6546	12310
$\|E\|$	120	4150	10630	37842
$\langle k \rangle$	0.307	1.957	3.248	6.148
k_{max}	4	18	31	47
std	0.581	2.095	3.145	5.195
cc	0.013	0.306	0.443	0.561
$avg(\|E_{cpnt}\|)$	0.179	2.262	5.502	21.141
$\|V_{giant}\|$	8	118	245	1239
$\|E_{giant}\|$	10	232	738	6393

Table 7: Network density

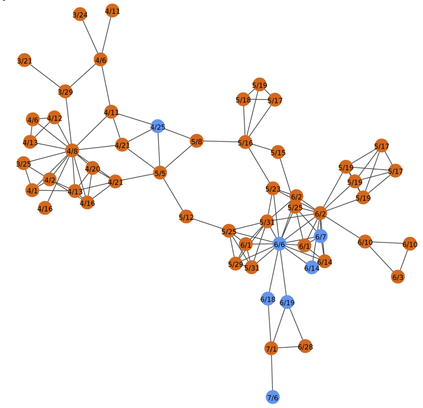
	G_{CDI}	$G_{RCDI3\%}$	$G_{RCDI5\%}$	$G_{RCDI10\%}$
$\frac{\|E\|}{\|E*\|}$	0.000392	0.000462	0.000496	0.000499
$\frac{\|E\|}{\|V\|}$	0.153257	0.978543	1.623892	3.074086
$\frac{\|E_{giant}\|}{\|V\|}$	0.012771	0.054704	0.112741	0.519334
$\frac{avg(\|E_{cpnt}\|)}{\|V\|}$	0.000229	0.000533	0.000841	0.001717

Contact Information

Email: {hankyu-jang, philip-polgreen, alberto-segre, daniel-sewell, sriram-pemmaraju}@uiowa.edu

Spatio-temporal clustering (Continued)

A connected component in GRCDI 10% composed of 7 CDI cases, in blue, and 48 asymptomatic C. diff carriers, in orange, over 3 months period (March 21 - July 6 2011). The CDI case (July 6) on the bottom-most of the graph is not connected to other CDI cases directly, but an asymptomatic C. diff carrier (July 1) connects them to CDI cases (June 18, June 19).



Discussion and future studies

Our results point to several avenues for future work that involve gathering prospective clinical data. Our Stage 1 model for identifying patients who are likely to be asymptomatic C. diff carriers needs to be clinically tested. Designing low-cost clinical protocols for gathering these data and performing appropriate statistical tests is critical in order to have confidence in our results. One of our findings suggests that risk factors for asymptomatic C. diff carriage include most of the standard risk factors, with the exception of high-risk antibiotics. This finding needs to be made more precise and also tested by gathering prospective clinical data.
In this paper, we only consider the possibility of CDI cases being exposed to asymptomatic C. diff carriers. We do not consider more complicated chains of exposure involving sequences of asymptomatic C. diff carriers. Combining more complicated exposure chains with individual risk models is another avenue for future work. It seems possible to use formulations that involve the Steiner tree problem for this purpose.
Another direction of the future work is using deep embedding approaches, such as Graph Convolutional Networks (GCN) where we let the deep neural network to learn from individual risk factors in the EMR and their exposure to other patients that are captured in the ADT data. Our asymptomatic C. diff carrier detection method can be applied in other infectious diseases where exposure plays an important role in disease diffusion. It is usually unknown if people we come in contact with are asymptomatic carriers of an infectious diseases. However, if data on an individual's risk factors to an infectious disease, contact information between these individuals, and a subset of individuals' infectious state is available, then our model would be able to detect the latent spreaders.