Data-driven Approaches to Risk Stratification and Asymptomatic Case Identification for HAIs

Data Seminar

03/19/2021

Healthcare-Associated Infections (HAIs)

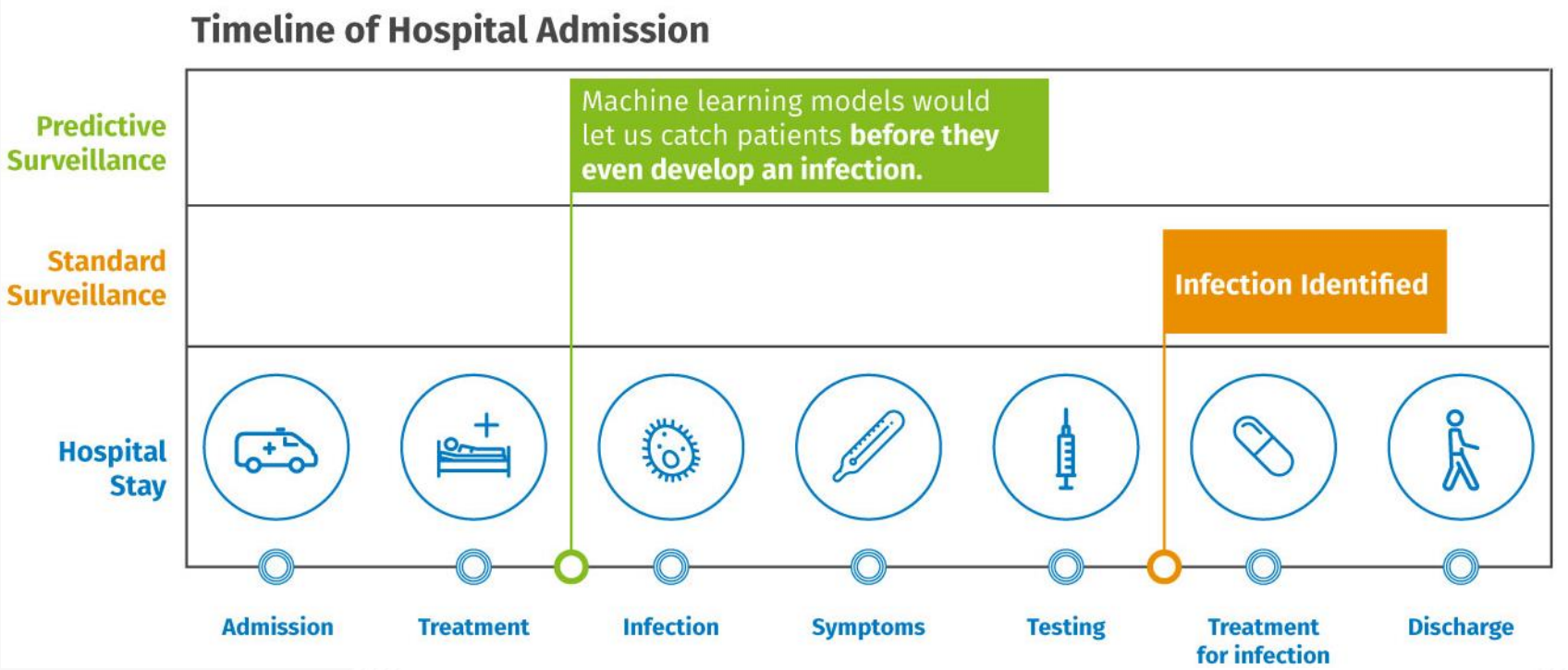


Figure from Wolters Kluwer

Invasive medical devices/procedures

- Catheter-associated urinary tract infections
- Ventilator-associated pneumonia
- Surgical-site infections
- And so on….

Clostridium Difficile (*C. diff*)

- Bacterial infection that attacks GI system.
- Transmitted by spores in patients' feces.
- Severe diarrhea, colitis, and mortality.
- 500,000 infections and 15,000 deaths annually in the US.
- No principled way of identifying asymptomatic patients.

With machine learning,

1. Can we ***predict patients' risk of infection***?
2. Can we ***detect asymptomatic spreaders***?

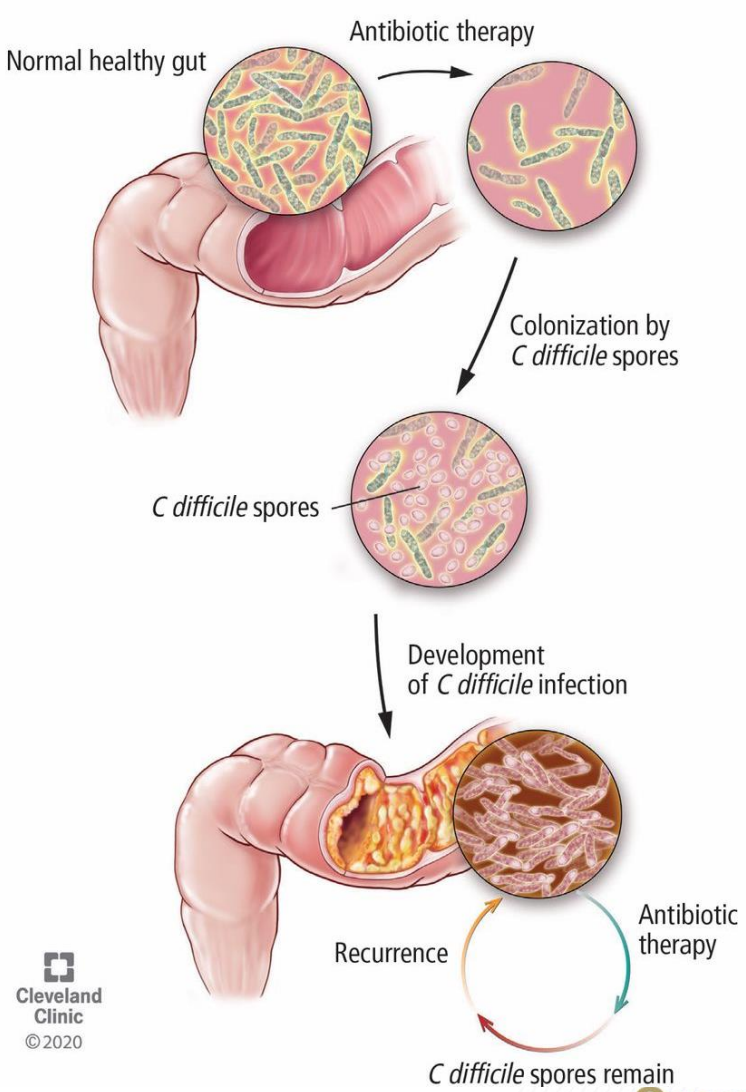


Figure from [Tsigrelis (2020)]

Papers

Learning the Probability of Activation in the Presence of Latent Spreaders

by Makar et al. (AAAI 2018)

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

by Jang et al. (epiDAMIK 2020)

Using Machine Learning and the Electronic Health Record to Predict Complicated Clostridium difficile Infection

by Li et al. (Open Forum Infect Dis. 2019)

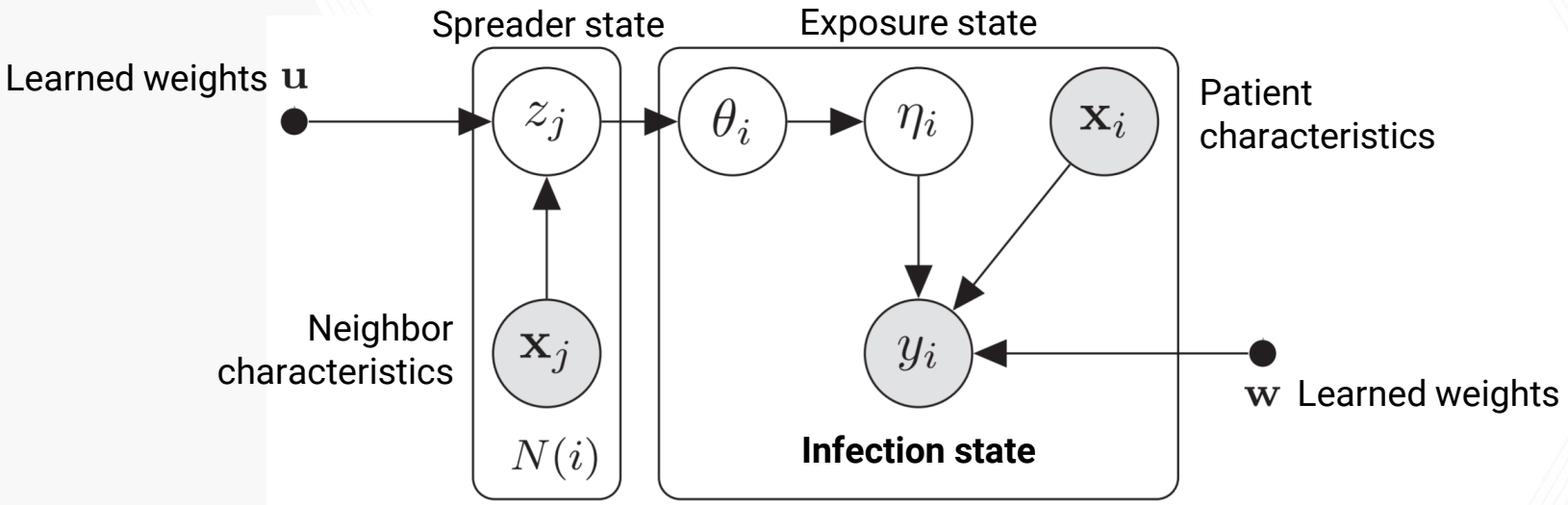
Contribution

Probability of Activation in the Presence of Latent Spreaders (PALS)

1. PALS can accurately estimate the risk of infection by modeling susceptibility and exposure.
2. The parameters in PALS lets us study varying significance of patient characteristics to infection and design interventions based on them.

Generative Model

For each patient i,



Inference

- E-step requires evaluating posterior distribution:

$$\frac{p(\boldsymbol{z}_i|\boldsymbol{u}, X_{n(i)})p(\theta_i|\boldsymbol{z}_i)p(\eta_i|\theta_i)p(y_i|\boldsymbol{x}_i, \eta_i, \boldsymbol{w})}{\int_\theta \sum_{\boldsymbol{z}} \sum_\eta p(\boldsymbol{z}_i|\boldsymbol{u}, X_{n(i)})p(\theta_i|\boldsymbol{z}_i)p(\eta_i|\theta_i)p(y_i|\boldsymbol{x}_i, \eta_i, \boldsymbol{w})}$$

- Exact inference is intractable due to $2^{|n(i)|+2}$ number of terms in denominator.

⇒ ***Variational inference***

$$q(z_i, \theta_i, \eta_i) = \prod_{j \in n(i)} q(z_j|\phi_j)q(\theta_i|\gamma_i)q(\eta_i|\pi_i)$$

ϕ_j ⇒ prob. of being a spreader
π_i ⇒ prob. of being exposed
γ_i ⇒ neighbors' influence

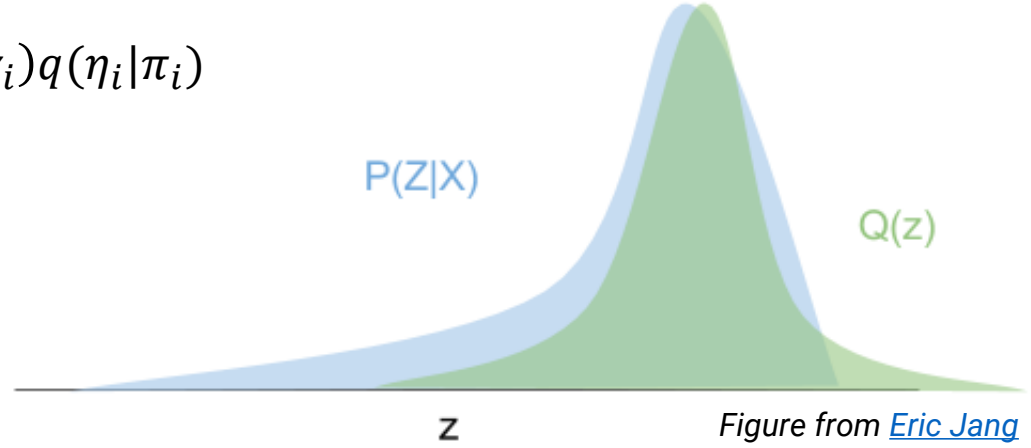


Figure from Eric Jang

Inference

E-step update of probability of patient j being a spreader (as neighbor of patient i):

$$\phi_{j,1} \propto \sigma(u^T x_j) \exp\left(\psi(\gamma_{j,1})\right)\left(1 + \sum_{k \neq j} \phi_{k,1}\right)^{-1}$$

Patient j's prob. of being a spreader

Patient i's prob. of being exposed

Patient i's other neighbors' spreader states

Patient i is exposed with high prob. but many neighbors other than j are spreaders.

$\Rightarrow$ Patient j is assigned small spreader probability.

Dataset

Objective: *Predict binary label indicating whether a patient was diagnosed with C. diff infection (CDI)* ***after the 5th day of hospitalization***.

1. **Study population**
   - Hospitalizations in large urban hospital from May 2012 to May 2014.
   - 350 cases of CDI out of 20,147 admissions. Temporal 50-50 train-test split.
2. **Contact networks**
   - *Nurse* network. Edge ⇔ drugs administered by the same nurse on same day.
   - *Room* network. Edge ⇔ spending any time during the same day in same room.
3. **Patient characteristics**
   - Demographics and previous medical history.
   - Ongoing procedures, medications, lab tests, location in hospital unit.
   - Up to day 5 as main patient *vs.* Up to date of contact as neighbor patient.

Results

- Baseline models: L1-regularized logistic regression.
- Nurse network gave better overall performance.
- *NoObs*: all spreaders are latent / *PartObs*: 10% are observed

Model	AUC (95% CI)
Susceptibility-only	0.698 (0.694, 0.703)
Susceptibility + Neighbor Infections	0.694 (0.693, 0.696)
PALS (NoObs)	0.700 (0.699, 0.702)
PALS (PartObs)	**0.705 (0.703, 0.706)**

Results

Weights in $\mathbf{u}$ (used for spreader state) from best-performing model

- Most negative weight in *"Receiving treatment for CDI"* feature.

  ⇒ Contact precautions are effective in hospitals.

- Most positive weights in *"Broad-spectrum antibiotics"/"Treatment for diarrhea"*.
  - General antibiotics are known to induce growth of *C. diff*.
  - Diarrhea increases the spread through use of restrooms.

Conclusion (Recap)

Probability of Activation in the Presence of Latent Spreaders (PALS)

1. PALS can accurately estimate the risk of infection by modeling susceptibility and exposure.
2. The parameters in PALS lets us study varying significance of patient characteristics to infection and design interventions based on them.

Papers

Learning the Probability of Activation in the Presence of Latent Spreaders

by Makar et al. (AAAI 2018)

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

by Jang et al. (epiDAMIK 2020)

Using Machine Learning and the Electronic Health Record to Predict Complicated Clostridium difficile Infection

by Li et al. (Open Forum Infect Dis. 2019)

Contribution

2-Stage classification model for asymptomatic carriers

1. 2-Stage model can predict asymptomatic *C. diff* carriers as well as indirectly validate results without “ground-truth” labels.
2. Exposure to asymptomatic carriers is a significant factor in determining the risk of CDI.

Dataset

1. Study population

- 154,230 patient visits in Univ. of Iowa hospitals from 2007 to 2011.
- After pre-processing, divided into $visit_{CDI}$ (750) and $visit_{CDIx}$ (115,271).
- Each $visit_{CDIx}$ generates one instance per day ⇒ 988,780 non-CDI instances.
- For $visit_{CDI}$, one per day until 3 days before diagnosis ⇒ 8,946 CDI instances.

2. Patient features

Baseline (B)

- Length of stay (LOS), age, gender, previous visits (PV).
- 5 high-risk antibiotics (ABX_i) and 2 gastric acid suppressors (GAS_i):

Antibiotics (ABX)

  - Binary prescription feature, Sum/Average prescription days.
- 4 exposure (patients are infectious 3 days before ~ 14 days after CDI result)
  - Cumulative/average daily number of CDI patients in same unit/room.

Colonization Pressure (CP)

Stage 1: Predict asymptomatic carriers

Hypothesis 1: Asymptomatic carriers and CDI cases have similar risk profiles.

⇒ Use CDI cases as predictive labels.

- Models based on Hypothesis 1: D^{B}, $D^{B,CP}$, $D^{B,ABX}$, $D^{B,ABX,CP}$
- 2-layer perceptron model with 80-20 train-test split.

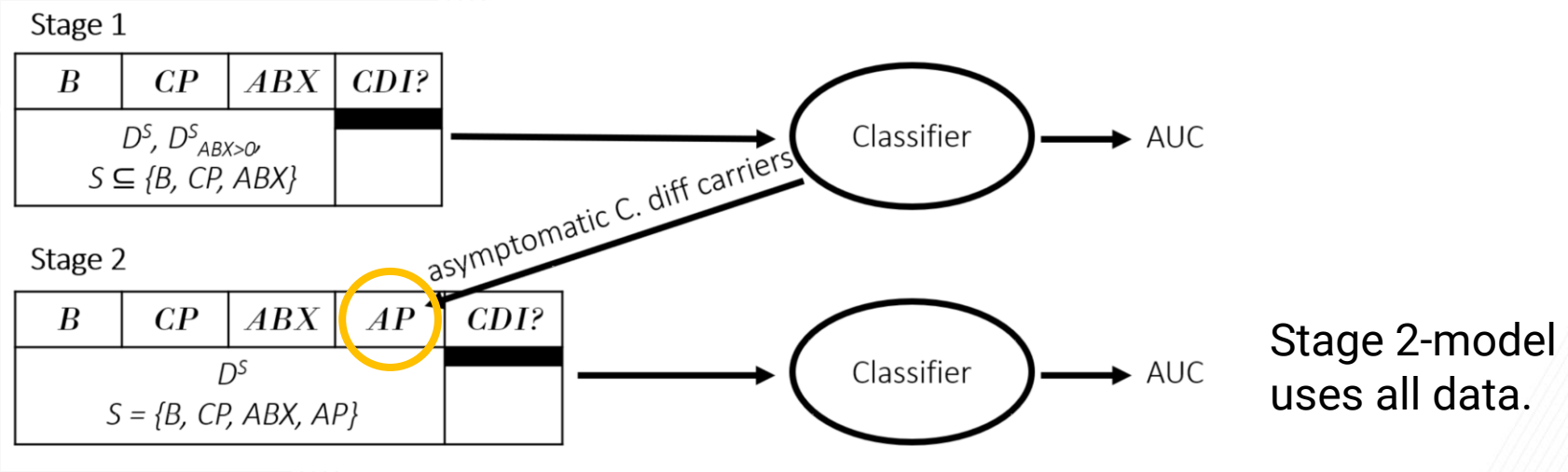
Hypothesis 2: The mechanism acquiring CDI consists of the patient first being an asymptomatic carrier and then being prescribed high-risk antibiotics.

⇒ Restrict dataset to patients prescribed at least one ABX.

- 5,483 CDI instances and 374,821 non-CDI instances.
- Models based on Hypothesis 2: $D^{B}_{ABX>0}$, $D^{B,CP}_{ABX>0}$, $D^{B,ABX}_{ABX>0}$, $D^{B,ABX,CP}_{ABX>0}$

Stage 2: Validate stage 1 models

- Each Stage 1-model returns the prob. of a patient being an asymptomatic carrier on that day.
- For each patient, take the maximum across all instances from the visit.
- Select top 10%, 5%, and 3% of visits in $visit_{CDIx}$ as asymptomatic carriers.



Stage 2-model uses all data.

Does the Stage 2 model perform better when including signals of exposure to asymptomatic carriers?

Results

Stage 1

- Using all standard risk-factors led to best performance.
- *ABX* and *CP* both help in finding CDI.
- *ABX*-restriction did not help.

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

Stage 2

- *ABX* is not associated with asymptomatic *C. diff* carriage.
- Exposure to asymptomatic C. diff carriers impacts the CDI spread.

AP	D^{B}	$D^{B,ABX}$	$D^{B,CP}$	$D^{B,CP,ABX}$
10%	0.712	0.687	**0.733**	0.700
5%	0.701	0.690	**0.727**	0.693
3%	0.689	0.698	**0.729**	0.710

Conclusion (Recap)

2-Stage classification model for asymptomatic carriers

1. 2-Stage model can predict asymptomatic *C. diff* carriers as well as indirectly validate results without “ground-truth” labels.
2. Exposure to asymptomatic carriers is a significant factor in determining the risk of CDI.

Papers

Learning the Probability of Activation in the Presence of Latent Spreaders

by Makar et al. (AAAI 2018)

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

by Jang et al. (epiDAMIK 2020)

Using Machine Learning and the Electronic Health Record to Predict Complicated Clostridium difficile Infection

by Li et al. (Open Forum Infect Dis. 2019)

Contribution

Electronic Health Record (EHR)-based predictive Model

1. EHR can accurately estimate risk of developing ***complicated CDI*** and outperforms models based on expert-curated features.
2. We can examine coefficients of the EHR model to interpret factors most associated with high or low risk of ***complicated CDI***.

Task

- Individual treatment of CDI is difficult.
- Genetic diversity requires careful selection of antibiotics (cost, resistance, etc.).

***Objective**: Given a patient has CDI, how likely is it that the infection becomes complicated?*

- Complicated CDI
    1. Admission to intensive care
    2. Toxic megacolon ⇒ Colectomy
    3. Mortality
- Predictions on the day of diagnosis,

  1 day after, or 2 days after.

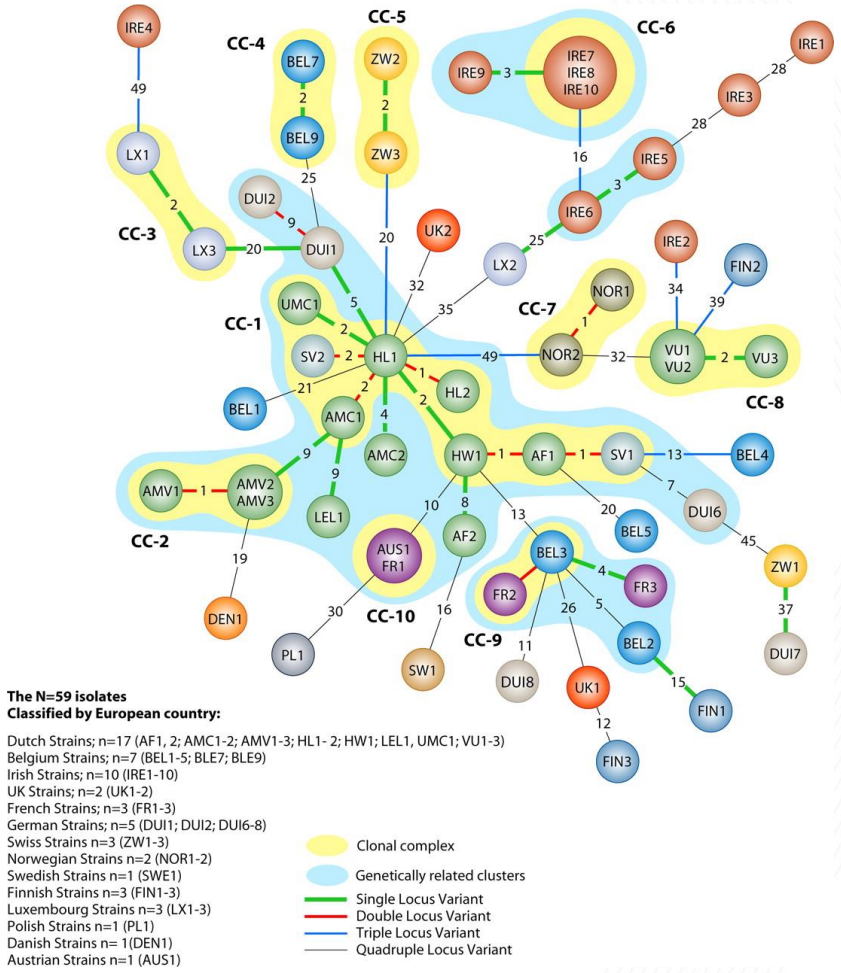
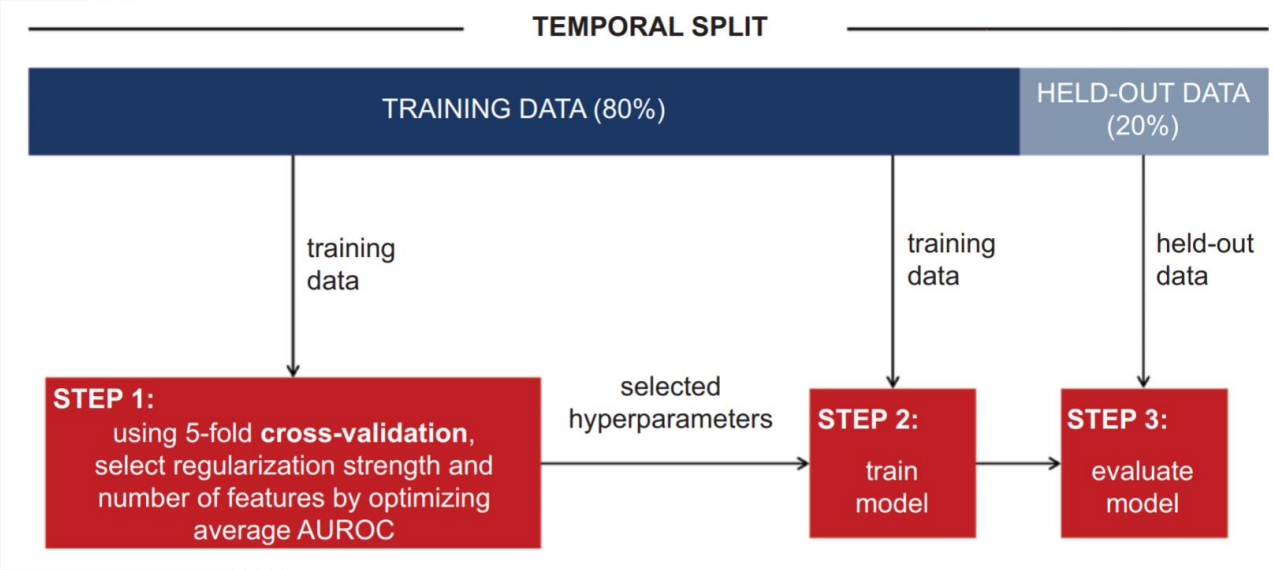


Figure from [Freeman et al.]

Dataset

1. **Study population**
   - 1118 CDI cases in Univ. of Michigan hospitals from October 2010 to January 2013.
   - 89 (8%) complicated CDI cases out of 1118.
2. **Complicated CDI labels**
   - Cases labeled through chart review by 2 clinicians. Viewed by 3rd if disagreed.
   - Cases labeled as complicated only if caused by CDI.
3. **Patient feature categories (# of features)**
   - *EHR (4271)*: Demographics and medical history of past 90 days from UM data repo.
   - *Curated (23)*: Expert-curated variables (e.g., age, cancer diagnosis) from Rao et al..

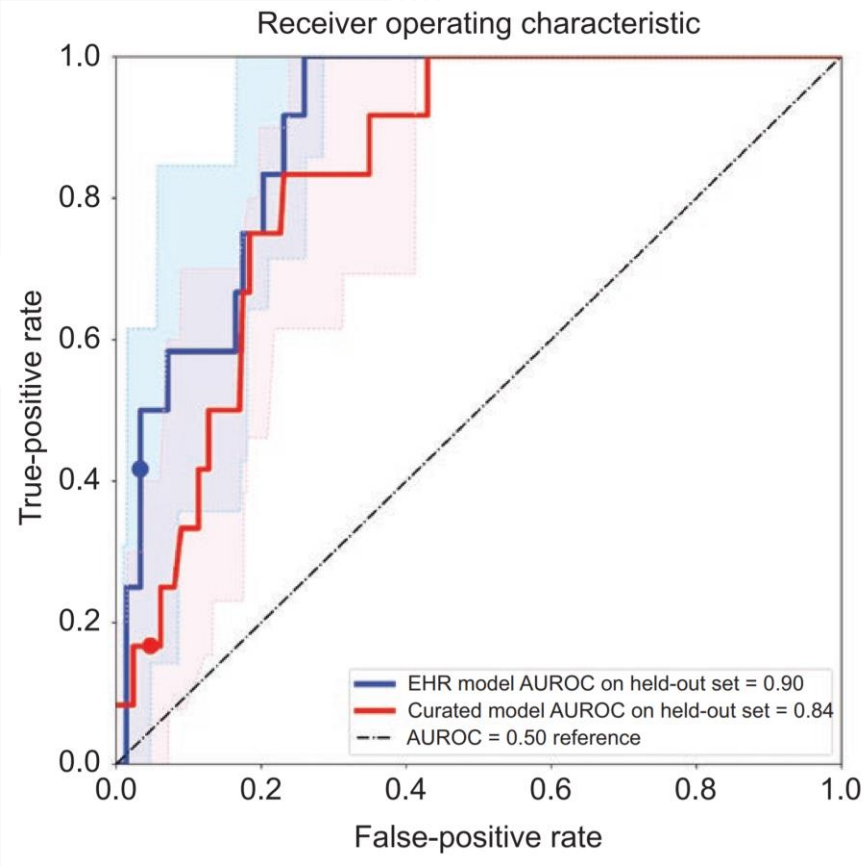
Model



- Logistic regression with L2-regularization and k-best feature selection
- Regularization parameter and number of features k picked with cross-validation.

Results

2 days after diagnosis, EHR outperforms Curated. (0.90 vs. 0.84)



EHR model

Actual label

Predicted label

TP 5	FP 7
FN 7	TN 205

Sens. = 41.7%
Spec. = 96.7%
PPV = 41.7%

Curated model

Actual label

Predicted label

TP 2	FP 10
FN 10	TN 202

Sens. = 16.7%
Spec. = 95.3%
PPV = 16.7%

Model (# of features)	AUROC (95% CI)
Curated (23)	0.84 (0.75-0.92)
EHR (900)	**0.90 (0.83-0.95)**
EHR+Curated (923)	0.88 (0.81-0.95)

Results

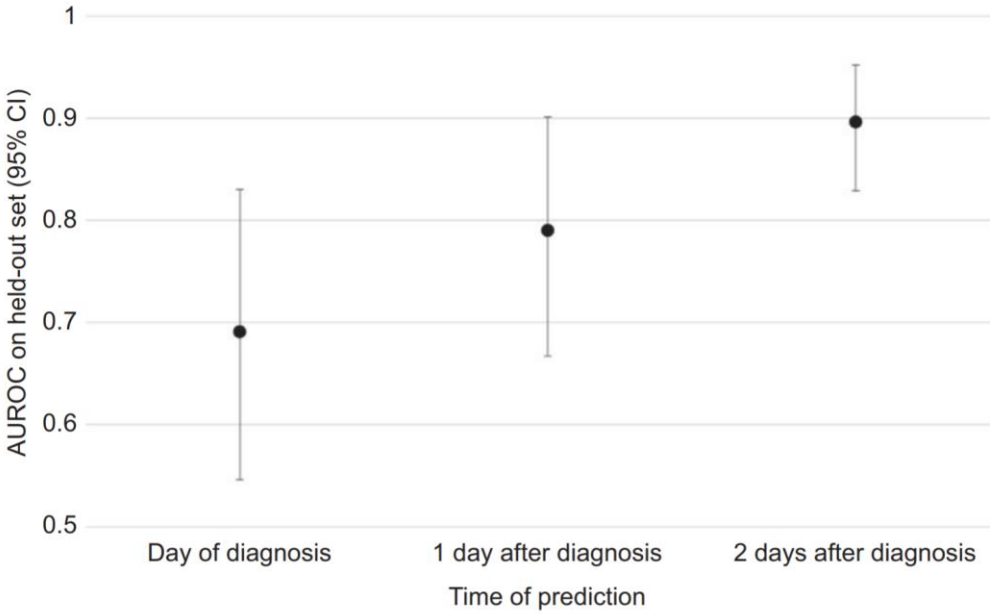
According to model coefficients, factors most associated

1) with risk

- High and low respiratory rates
- Low systolic blood pressure
- Low blood CO_2

2) with protection

- Normal respiratory rate
- Young age



Model performance decreases when making predictions earlier.

Conclusion (Recap)

Electronic Health Record (EHR)-based predictive Model

1. EHR can accurately estimate risk of developing complicated CDI and outperforms models based on expert-curated features.
2. We can examine coefficients of the EHR model to interpret factors most associated with high or low risk of complicated CDI.

References

“Learning the probability of activation in the presence of latent spreaders”, Makar M, Guttag J, Wiens J. Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).

“A Data-driven Approach to Identifying Asymptomatic C. diff Cases”, Jang H, Polgreen P M, Segre A M, et al. 2020.

“Using machine learning and the electronic health record to predict complicated Clostridium difficile infection”, Li B Y, Oh J, Young V B, et al. Open forum infectious diseases. US: Oxford University Press, 2019, 6(5): ofz186.

“Recurrent Clostridioides difficile infection: Recognition, management, prevention”, Constantine Tsigrelis, Cleveland Clinic Journal of Medicine Jun 2020, 87 (6) 347-359; DOI: 10.3949/ccjm.87gr.20001

“Predicting hospital infections: how AI makes it possible”, Wolters Kluwer, https://www.wolterskluwer.com/en/expert-insights/predicting-hospital-infections-how-ai-makes-it-possible, Date accessed: March 19, 2021

“A Beginner's Guide to Variational Methods: Mean-Field Approximation”, Jang E https://blog.evjang.com/2016/08/variational-bayes.html, Date accessed: March 19, 2021

“The Changing Epidemiology of Clostridium difficile Infections” J. Freeman, M. P. Bauer, S. D. Baines, J. Corver, W. N. Fawley, B. Goorhuis, E. J. Kuijper, M. H. Wilcox, Clinical Microbiology Reviews Jul 2010, 23 (3) 529-549; DOI: 10.1128/CMR.00082-09