Risk Stratification for Health-Care Associated *C. diff*

Learning Evolving Patient Risk Processes for *C. diff* Colonization

**Jenna Wiens**, John V. Guttag and Eric Horvitz*

Massachusetts Institute of Technology

* Microsoft Research
**Clostridium difficile (C. diff)**

- Bacteria takes over the gut when normal flora gets wiped out
- Transmitted through the mouth
- Causes severe diarrhea, intestinal diseases
- Treatment: metronidazole, oral vancomycin
- 20% of cases relapse within 60-days

(Pepin J et al., 2005)

[http://newsimg.bbc.co.uk/media/images/41223000/jpg/_41223637_clostridium203spl.jpg](http://newsimg.bbc.co.uk/media/images/41223000/jpg/_41223637_clostridium203spl.jpg)
Prevalence

- Hospital-acquired: 178,000/year
  (McDonald et al., 2006)
- On par with number of new cases of invasive breast cancer in the US each year
  (American Cancer Society, 2009)

(CDC, 2011)
Risk Factors

**Time Invariant**
- Collected at the time of admission
- *e.g.*, admission complaint, previous admissions, home meds

**Time Varying**
- Changes during the hospitalization
- *e.g.*, current meds, current procedures, current location, hospital conditions

Representing and reasoning about temporal processes promises to enhance the accuracy of inferences about risk.
Typical Approach in Clinical Literature

Risk based on patient’s state at time of admission
(Tanner et al., 2009)

Index Event

Risk based on patient’s state \( x \) days before index event
(Dubberke et al., 2011)
Our Approach

Estimate Risk Profile

Time (days)

P(Index Event)

Index Event
Risk Processes

Hypothesis: extracting and analyzing evolving patient risk can lead to a more accurate model for predicting infections

Risk Process: describes the evolution of risk over the course of a hospital admission
Inferring Risk Processes

• Challenges:
  • No ground truth about risk
    o Retrospective data $\rightarrow$ not all patients get tested
    o Actual risk on any day is unknowable
  • Thousands of correlated variables
The Data

- Database from a large urban hospital in the US
- In-patient stays from a single year
- Inclusion criteria (see paper for details)
  - Eliminate easily identifiable cases

Population:
- ~10,000 hospital admissions
- ~200 Positive *C. diff* cases
Experimental Setup

- Training & Testing
  - Randomly subsampled the negative class
  - Split data into stratified training and test sets 70/30.
  - Training set 1,251 admissions (127 positive)
  - Testing set 532 admissions (50 positive)
# Features

<table>
<thead>
<tr>
<th>Time Invariant</th>
<th>Time Varying</th>
</tr>
</thead>
<tbody>
<tr>
<td>prev. ICD 9 codes</td>
<td>lab results</td>
</tr>
<tr>
<td>home medications</td>
<td>procedures</td>
</tr>
<tr>
<td>prev. admission medications</td>
<td>location room</td>
</tr>
<tr>
<td>patient’s city</td>
<td>location unit</td>
</tr>
<tr>
<td>attending MD</td>
<td>medications</td>
</tr>
<tr>
<td>Hospital service</td>
<td>vitals</td>
</tr>
<tr>
<td>admission source</td>
<td>day of admission</td>
</tr>
<tr>
<td>financial class code</td>
<td>unit CP</td>
</tr>
<tr>
<td>admission complaint</td>
<td>hospital wide CP</td>
</tr>
<tr>
<td>admission procedure</td>
<td></td>
</tr>
<tr>
<td>patient’s race</td>
<td></td>
</tr>
</tbody>
</table>
Features: >10,000 variables for each day of every hospital admission

<table>
<thead>
<tr>
<th>Time Invariant</th>
<th>Time Varying</th>
</tr>
</thead>
<tbody>
<tr>
<td>• prev. ICD 9 codes</td>
<td>• lab results</td>
</tr>
<tr>
<td>• home medications</td>
<td>• procedures</td>
</tr>
<tr>
<td>• prev. admission medications</td>
<td>• location room</td>
</tr>
<tr>
<td>• patient’s city</td>
<td>• location unit</td>
</tr>
<tr>
<td>• attending MD</td>
<td>• medications</td>
</tr>
<tr>
<td>• Hospital service</td>
<td>• vitals</td>
</tr>
<tr>
<td>• admission source</td>
<td>• day of admission</td>
</tr>
<tr>
<td>• financial class code</td>
<td>• unit CP</td>
</tr>
<tr>
<td>• admission complaint</td>
<td>• hospital wide CP</td>
</tr>
<tr>
<td>• admission procedure</td>
<td></td>
</tr>
<tr>
<td>• patient’s race</td>
<td></td>
</tr>
</tbody>
</table>

- patient’s age
- patient’s marital status
- patient’s sex
- expected surgery
- ER admission
- dialysis
- diabetic
- history of C. diff
- num. hospital visits (90 days)
- avg., max., total los (90 days)
Representing a hospital stay:
Our Approach to Risk Stratification

\[
p_1 p_2 \ldots p_n \rightarrow \text{Features} \rightarrow \text{SVM} \rightarrow \vec{r}_1 \rightarrow \vec{r}_2 \rightarrow \ldots \rightarrow \vec{r}_n \rightarrow \overline{r} = [r_1, r_2, \ldots, r_n] \rightarrow \text{Risk Stratify} \rightarrow y
\]
Our Approach to Risk Stratification

SVM requires labels.
Labeling the Data

Ground Truth

Index Event

Risk

Days from Admission
Labeling the Data

- Positive Test
- Ground Truth
- Risk
- Days from Admission
Labeling the Data

Positive Test

Days from Admission

\( \chi \)
Labeling the Data

We assign each day of admission in which a patient eventually tests positive as positive (high risk)...

...and negative (low risk) otherwise.

We hope to identify high risk patients as early as possible.
Learning the Decision Boundary

We expect a patient’s risk fluctuates → noise in the training labels
Learning the Decision Boundary

Note: Simplified illustration. We learn a linear hyperplane in the high dimensional feature space.
Daily Risk -> SVM Continuous Predictions

We consider the distance each feature vector lies from the SVM decision boundary this results in a continuous prediction for each day.

\[ r_d = w \cdot p_d - b \]
Our Approach to Risk Stratification

\[ p_1, p_2, \ldots, p_n \]

\[ r = [r_1, r_2, \ldots, r_n] \]

SVM

Time Invariant

Time Varying

Risk Stratify

\[ y \]
Example Risk Processes
Using Risk Processes for Risk Stratification

• Instantaneous approach:
  – Analogous to typical risk stratification approaches
  – Considers value of risk process only on day of prediction

• Cumulative approach:
  – Combine estimates from all previous days
    E.g., constant, linear, and quadratic weighted averages

(Dubberke et al., 2011)
Evaluating Instantaneous Approach

- Consider **instantaneous** estimate for patient risk at a constant distance before the index event e.g., 2 days

Compute classifier performance by sweeping the decision threshold from min to max.
Evaluating Cumulative Approach

- **Combine** estimates for patient risk from the time of admission up to a constant distance from the index event e.g., 2 days

Patient tests positive for *C. diff* on day 8

\[ \text{PatientRisk}' = f([r_1, r_2, \ldots, r_{\text{index}-2}]) \]

Compute classifier performance by sweeping the decision threshold from min to max.
Defining the Index Event

• Positive Examples ➔ day of positive test result
  – We consider only data collected up to two days before a positive test result

• Negative Examples ➔ midpoint of admission
  – Considering discharge as the index event can lead to deceptively good results
## Results

<table>
<thead>
<tr>
<th>Approach</th>
<th>Testing AUROC (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant weighted avg.</td>
<td>0.7518 (0.69-0.81)</td>
</tr>
<tr>
<td>Linear weighted avg.</td>
<td>0.7444 (0.67-0.80)</td>
</tr>
<tr>
<td>Quadratic weighted avg.</td>
<td>0.7360 (0.67-0.80)</td>
</tr>
<tr>
<td>Instantaneous</td>
<td>0.6870 (0.61-0.77)</td>
</tr>
</tbody>
</table>
Results

Patients in the 5th quintile are at >20-fold greater risk than those in the 1st quintile!
Conclusion

• First step in analyzing how patient risk for acquiring *C. diff* may evolve during a hospitalization
  – Improvement over existing methods

• Next steps:
  – Find patterns of risk that lead to worse/better outcomes
  – Investigate application in other contexts (e.g., other HAIs, in-hospital mortality, LOS)
Acknowledgements

• Natural Sciences and Engineering Research Council of Canada (NSERC)
• National Science Foundation (NSF)
• Quanta Computer Inc.
Works Cited


