Improving risk prediction of Clostridium Difficile Infection using temporal event-pairs

Mauricio Monsalve,
Sriram Pemmaraju, Sarah Johnson, Philip Polgreen

Computational Epidemiology (compepi) group
The University of Iowa
Clostridium Difficile Infection (CDI) is a contagious HAI that burdens healthcare and is becoming increasingly deadly

We improve CDI prediction using of an ensemble of logistic regression classifiers, that processes patient visits described as pairs of events (chronologically orders)

Extensive feature selection to prevent overfitting

We apply our approach to a rich dataset from the University of Iowa Hospitals and Clinics (UIHC)

We produce better risk predictions (AUC) than existing estimators and identify novel risk factors.
1. At a glance
2. Clinical motivation
3. Data mining motivation
4. Proposed method
5. Results
6. Concluding remarks
Clinical Motivation

- In the United States, during 2011 alone
  - half a million patients suffered from CDI
  - 29,000 died within 30 after diagnosis
- CDI is specially troublesome because
  - threatens the weakest patients
  - is triggered by antibiotics of choice
  - survives alcohol, reduced gastric acid, and dryness of environment (spores, for months)
  - costly: extra days, expensive antibiotics
*Clostridium difficile* (*C. difficile*) causes life-threatening diarrhea. These infections mostly occur in people who have had both recent medical care and antibiotics. Often, *C. difficile* infections occur in hospitalized or recently hospitalized patients.

**Resistance of Concern**

- Although resistance to the antibiotics used to treat *C. difficile* infections is not yet a problem, the bacteria spreads rapidly because it is naturally resistant to many drugs used to treat other infections.
- In 2000, a stronger strain of the bacteria emerged. This strain is resistant to fluoroquinolone antibiotics, which are commonly used to treat other infections.
- This strain has spread throughout North America and Europe, infecting and killing more people wherever it spreads.

**Public Health Threat**

- 250,000 infections per year requiring hospitalization or affecting already hospitalized patients.
- 14,000 deaths per year.
- At least $1 billion in excess medical costs per year.
- Deaths related to *C. difficile* increased 400% between 2000 and 2007, in part because of a stronger bacteria strain that emerged.
- Almost half of infections occur in people younger than 65, but more than 90% of deaths occur in people 65 and older.
- About half of *C. difficile* infections first show symptoms in hospitalized or recently hospitalized patients, and half first show symptoms in nursing home patients or in people recently cared for in doctors’ offices and clinics.
• Clinical motivation:
  ◦ To help in the early identification of patients at high risk of developing CDI
• Why?
  ◦ Prepare to treat a patient for CDI
  ◦ Preventive isolation (minimize spread)
  ◦ Targeted sanitization
  ◦ Observe nearby patients
• Early identification?
  ◦ Best use patient's visit history to assess risk
**Data Mining Motivation**

- **Architecture**
  - ~19,600 rooms
  - ~570 bedrooms

- **ADT**
  - ~120,000 transfers
  - Since 2005

- **Quality**
  - ~200,000 admissions
  - ~120,000 patients
  - ~7,800,000 prescriptions
  - Since 2007

- **CDI tests**
  - ~2,000 positive tests results
  - Since 2005

- **EMR**
  - ~19,800,000 logins
  - ~17,500 computers
  - From 2006-09 Until 2008-06

**ICD9 codes**
- ~14,600 diagnosis codes
- ~3,900 procedure codes
To estimate the risk of a patient of developing CDI by using the data on the patient's visit so far

- Order of clinical events relevant to onset of CDI
- To describe a patient's visit as ordered events
• **Difficulties:**

  ◦ CDI affected patients are a **minority:** ~2,000 v. ~200,000
  ◦ CDI patients arrived diseased or left early (<1,000)
  ◦ **Sparsity** of events: a patient can only be associated to very, very few diagnoses, procedures, prescriptions
  ◦ **Feature explosion:** combinations of clinical events generate too many features (millions with just two events)
  ◦ Summing up: **computational cost** + risk of **overfitting**
**Proposed Method**

- To describe visits using **pairs of events**
- To rely on an **ensemble**
  - Counter **class imbalance**
  - **Split computational cost**
- **Logistic regression model** in each unit of the ensemble
  - Remove irrelevent **features**, while
  - **Minimize BIC** to quasi-maximize out-of-sample validity
  - Using **regularization**
Chronologically ordered pairs of events

- Only pairs of events?
  - Partial orders of minimal complexity
  - *In principle*, induce millions of features

- (x,y) or “[x < y]” reads as
  - Event x occurred before event y
  - Or, both events occurred in the same day

- Examples:
  - [To=OR < To=MICU]
  - [Proc=216 < RxMin=812]
  - [@Diag=135 < @Age=50]
• Admission data is treated as events
• Examples:
  ○ @Age=20
  ○ @Severity=HIGH
  ○ @Diag=135
  ○ @DiagPrev=135
• Manufactured events
  ○ @pcr_period
  ○ @cdi_1year
  ○ Pressure=HIGH
Hierarchies

• Available hierarchies:
  ◦ Medications, procedures, diagnoses

• Hierarchies can be revealing:
  ◦ E.g., are particular antibiotics risk factors or the whole category of antibiotics is a risk factor?

• How to consider hierarchies?
  ◦ Let \((x,y)\) be a pair of events, and \(x:S\) and \(y:T\)
  ◦ Besides \((x,y)\), consider also \((S,y)\), \((x,T)\), \((S,T)\)

• If we plan to prune features thoroughly, might as well introduce tentative features
• Individual classifiers: **logistic regression**

• Why?
  ○ **Binary features**—logistic regression is MaxEnt
  ○ **Sparsity linearizes** associations
  ○ **Regularization** possible
  ○ **Sparsity + L1 regularization**—L1-L0 equivalence

• Feature selection is cheap[er] with regularization

• Fast feature selection scheme—**two passes**
  ○ **First pass**: fast, inaccurate—remove low impact features
  ○ **Second pass**: slow[er], accurate—L1 regularization
Features organized into buckets

\[ b_1 \rightarrow \text{Top } b \text{ features} \rightarrow b_2 \rightarrow \text{Top } b \text{ features} \rightarrow b_3 \rightarrow \text{Top } b \text{ features} \rightarrow b_4 \rightarrow \cdots \rightarrow b_n \rightarrow \text{Top } b \text{ features} \rightarrow \text{Candidate features for BIC minimization} \]
**Algorithm 1** Greedy randomized embedded feature filter

**Input:** \( m \): bucket size, \( \mathcal{H} \): features, \( \mathcal{S} \): rows of the data set

1. Randomly partition \( \mathcal{H} \) into sets \( B_1, \ldots, B_k \), so that \( |B_i| = m \) for every \( i \) such that \( 1 \leq i \leq k - 1 \).
2. Let \( C = B_1 \).
3. **for** \( i = 2 \) to \( k \) **do**
4. \hspace{1em} Fit logistic regression model to \( \mathcal{S} \) projected on \( C \cup B_i \)
5. \hspace{1em} For \( h \in C \cup B_i \), define \( s(h) \) as the number of classification errors introduced when \( \beta_h \) is set to 0 (\( \beta_h \) is the coefficient of the logistic regression model for feature \( h \)).
6. \hspace{1em} Update \( C \) to be the \( m \) features in \( C \cup B_i \) with the highest \( s(h) \).
7. **end for**
8. **return** \( C \)
• Step 2: minimize BIC defined as

\[ BIC = -2L + (1 + |\beta|_0) \ln |S|, \]

where \( L \) is defined as

\[ L(\alpha, \beta; \lambda) = \lambda |\beta|_1 + \sum_{(x, y) \in S} \ln \left( 1 + \exp(-y(\alpha + \beta^T x)) \right), \]

by searching using \( \alpha, \beta, \lambda \)

• Encouraged by L0–L1 equivalence
RESULTS

Several experiments

1. Using only two days worth of data (comparison against state of the art: Wiens et al 2014)—85% v. 80% accuracy

2. Using more days worth of data—using pairs of events v. bare events: 86% v. 85%

3. What occurs to risk estimate as onset of CDI nears—sensitivity increases

4. Admission data v. strictly clinical events—83% v. 79%

5. Impact of BIC minimization step—+2% out-of-sample accuracy and 1,500 features removed
Experiment 1 (v. s-o-a)  Experiment 2 (pairs v. bare)
Experiment 3: risk curves (sensitivity)
Experiment 4:
Admission data versus clinical events only
• Admission data very predictive
• Required for prediction
• Clinical events only limited predictive ability
<table>
<thead>
<tr>
<th>Classifier</th>
<th>AUC</th>
<th>Active features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>With</td>
<td>Without</td>
</tr>
<tr>
<td>1-2 days</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BEC</td>
<td>85.07%</td>
<td>84.10%</td>
</tr>
<tr>
<td>PEC</td>
<td>86.20%</td>
<td>83.97%</td>
</tr>
<tr>
<td>Any day</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BEC</td>
<td>85.26%</td>
<td>84.25%</td>
</tr>
<tr>
<td>PEC</td>
<td>86.61%</td>
<td>84.49%</td>
</tr>
<tr>
<td>Later days</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BEC</td>
<td>85.21%</td>
<td>84.10%</td>
</tr>
<tr>
<td>PEC</td>
<td>86.53%</td>
<td>84.34%</td>
</tr>
</tbody>
</table>

Table VI: Impact of $L_1$-regularization with BIC minimization on the classifiers. For each classifier, the AUC on the any day, later days and 1-2 days testing sets are presented, as well as the number of features, for the cases with and without regularization. The values are averaged over the 10 fold cross validation tests.
<table>
<thead>
<tr>
<th>Feature name</th>
<th>Log odds ($\beta_i$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>@Diag=135</td>
<td>5.2718</td>
</tr>
<tr>
<td>@Severity=Major</td>
<td>2.7682</td>
</tr>
<tr>
<td>@Severity=Extreme</td>
<td>2.4677</td>
</tr>
<tr>
<td>@Severity=Moderate</td>
<td>1.8935</td>
</tr>
<tr>
<td>@AdmSrc=NEONOBORN PREMATURITY BIRTH</td>
<td>1.7402</td>
</tr>
<tr>
<td>[@Severity=Minor &lt; @AdmType=ELECTIVE/ROUTINE]</td>
<td>1.4477</td>
</tr>
<tr>
<td>@SvcCat=INTERNAL MEDICINE</td>
<td>1.1968</td>
</tr>
<tr>
<td>@Diag=203</td>
<td>-1.1257</td>
</tr>
<tr>
<td>@AGE=20</td>
<td>-1.1054</td>
</tr>
<tr>
<td>[To=PORR &lt; To=OR]</td>
<td>-1.0547</td>
</tr>
<tr>
<td>@PCR_period</td>
<td>-0.9199</td>
</tr>
<tr>
<td>[@PCR_period &lt; @Diag=152]</td>
<td>0.8856</td>
</tr>
<tr>
<td>[@SvcCat=INTERNAL MEDICINE &lt; @AdmType=ELECTIVE/ROUTINE]</td>
<td>0.8567</td>
</tr>
<tr>
<td>@SvcCat=FAMILY MEDICINE</td>
<td>-0.8476</td>
</tr>
<tr>
<td>[@SvcCat=PSYCHIATRY &lt; @AdmType=EMERGENCY]</td>
<td>-0.8273</td>
</tr>
<tr>
<td>@AGE=30</td>
<td>-0.7447</td>
</tr>
<tr>
<td>[@AdmType=URGENT &lt; @AGE=20]</td>
<td>-0.7261</td>
</tr>
<tr>
<td>@AdmSrc=UIHC CLINIC</td>
<td>-0.7136</td>
</tr>
<tr>
<td>@Diag=52</td>
<td>0.6986</td>
</tr>
<tr>
<td>@Readm_90D</td>
<td>0.6706</td>
</tr>
</tbody>
</table>

Table VIII: Top 20 most influential features in PEC, any day.
<table>
<thead>
<tr>
<th>Feature name</th>
<th>$\beta_{x&lt;y}$</th>
<th>$\beta_{y&lt;x}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>[To=OR $&lt;$ To=PORR]</td>
<td>-0.1831</td>
<td>-1.0547</td>
</tr>
<tr>
<td>transferred to OR no later than transferred to PORR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[To=OR $&lt;$ Proc=Diag/Therap]</td>
<td>0.3556</td>
<td>-0.0004</td>
</tr>
<tr>
<td>transferred to OR no later than underwent miscellaneous diagnostic-therapeutic procedure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Proc=223 $&lt;$ Proc=231]</td>
<td>0.1982</td>
<td>-0.0019</td>
</tr>
<tr>
<td>underwent 'enteral and parenteral nutrition' no later than underwent 'other therapeutic procedures'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Proc=231 $&lt;$ RxMin=812]</td>
<td>-0.014</td>
<td>-0.2062</td>
</tr>
<tr>
<td>underwent 'other therapeutic procedures' no later than prescribed 'antibiotics systemic'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Proc=Diag/Therap $&lt;$ RxSmin=81206]</td>
<td>0.1781</td>
<td>0.0144</td>
</tr>
<tr>
<td>underwent miscellaneous diagnostic-therapeutic procedure no later than prescribed 'fourth generation cephosphorins'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[RxMaj=40 $&lt;$ RxSmin=562210]</td>
<td>0.1362</td>
<td>0.0079</td>
</tr>
<tr>
<td>prescribed 'nutrients/nutritional agents' no later than prescribed '5ht3 receptor antagonists'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Proc=216 $&lt;$ RxSmin=81219]</td>
<td>0.0042</td>
<td>-0.1182</td>
</tr>
<tr>
<td>underwent 'respiratory intubation and mechanical ventilation' no later than prescribed 'extended-spectrum penicillin'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[To=6RCE $&lt;$ Proc=Diag/Therap]</td>
<td>-0.0055</td>
<td>-0.112</td>
</tr>
<tr>
<td>transferred to 6RCE no later than underwent miscellaneous diagnostic-therapeutic procedure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[RxSmin=280892 $&lt;$ RxSmin=81203]</td>
<td>0.1132</td>
<td>0.0096</td>
</tr>
<tr>
<td>prescribed 'misc analgesics systemic' no later than prescribed 'first generation cephosphorins'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Proc=177 $&lt;$ RxSmin=280808]</td>
<td>-0.0161</td>
<td>-0.1191</td>
</tr>
<tr>
<td>underwent 'computerized axial tomography (ct) scan head' no later than prescribed 'opiate agonists'</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.7. Top 10 ordered events where the order is relevant. For each ordered event, we include a human readable description of it as well as two log-odds: the individual log-odds of the ordered pairs $\beta_{x<y}$, and its converse $\beta_{y<x}$.
CONCLUDING REMARKS

• Possible to outperform literature, but admission data is very predictive
  ◦ Event data introduces marginal improvements
  ◦ Useful for risk curves—impossible with admission data

• CDI Colonization Pressure deemed irrelevant by the classifier

• Future work
  ◦ Improvement of classification methodology
    ▪ Better distinguish relevant features (order, hierarc.)
    ▪ Trade-off size of ensemble, complexity of units
  ◦ Further study role of transmission in CDI