Identifying Clostridium Difficile in the ICU Using Bayesian Networks

Phenome Based Analysis

Peijin Zhang
Second Annual MIT PRIMES Conference

May 20th, 2012

Introduction

- Development of comprehensive electronic medical record databases has made large phenome wide association studies possible
- Prior work has tried to link phenomes with molecular information, such as SNPs
- Feasibility of these studies on the Mimic II database has been shown through association studies with other variables, such as required fluid levels
- Our project strives to link phenomes defined by ICD9 codes with lab values of patients in the ICU.

MIMIC II

- Database with 30,000 ICU patients between 2001 and 2007.
- Comprehensive data: Lab tests, International Classification of Diseases (ICD9) discharge codes, medication orders, etc.
- 5675 distinct ICD9 codes (33.5 % of all possible)
- Mean of 8.7 ICD9 codes assigned to each admission.
- Allows for large scale phenome based analysis

White Blood Cell Count (WBC) Use Case

- Mimic stores all measured lab events of patients (748 total)
- Use case focuses on WBC, commonly measured lab event present in most patients.

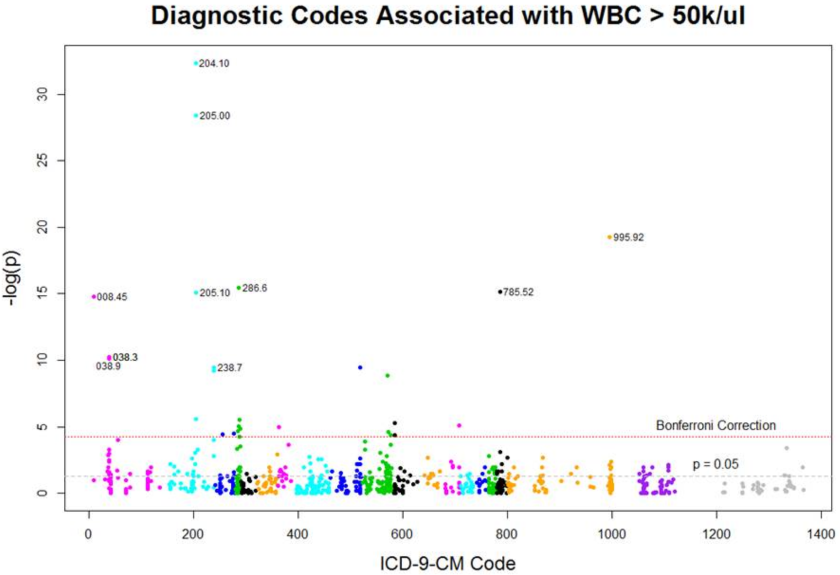
White Blood Cell Count (WBC) Use Case

- Mimic stores all measured lab events of patients (748 total)
- Use case focuses on WBC, commonly measured lab event present in most patients.
- Method:
    - Single WBC lab result chosen for each patient by taking peak value during ICU stay
    - Patient subsets chosen by creating lower bounds from a hundred equally spaced cutoffs between counts of 0 and 100,000/μL (IE $\geq$ 20,000/μL, $\geq$ 80,000/μL, etc)
    - Exact binomial test was performed on each subset for occurence of each ICD9 code with $\geq$ 100 disctinct cases when compared to the full database.

White Blood Cell Count (WBC) Use Case

- Mimic stores all measured lab events of patients (748 total)
- Use case focuses on WBC, commonly measured lab event present in most patients.
- Method:
  - Single WBC lab result chosen for each patient by taking peak value during ICU stay
  - Patient subsets chosen by creating lower bounds from a hundred equally spaced cutoffs between counts of 0 and 100,000/μL (IE $\geq$ 20,000/μL, $\geq$ 80,000/μL, etc)
  - Exact binomial test was performed on each subset for occurence of each ICD9 code with $\geq$ 100 disctinct cases when compared to the full database.
  - Results filtered using Bonferroni Correction (p value less than $\frac{0.05}{\text{number of ICD9 codes}}$)

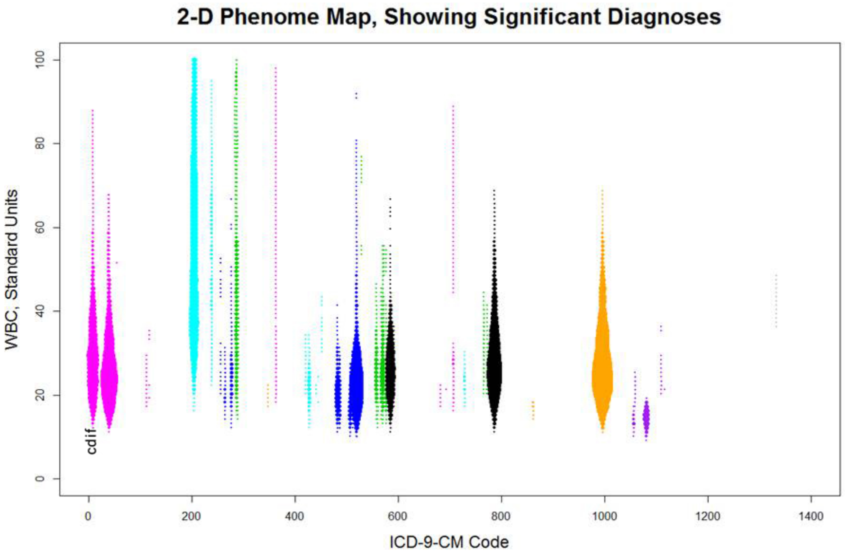
Single Lower Bounded Run



Significant ICD9 Codes

ICD9 Code	Description	P value
204.10	Chronic lymphoid leukemia (CLL)	4.3×10^{-33}
205.00	Acute myeloid leukemia (AML)	4.0×10^{-29}
995.92	Severe sepsis	5.3×10^{-20}
286.6	Disseminated intravascular coagulation (DIC)	3.7×10^{-16}
785.52	Septic shock	7.3×10^{-16}
205.10	Chronic myeloid leukemia (CML)	7.7×10^{-16}
008.45	**Intestinal infection due to Clostridium difficile**	1.7×10^{-15}
038.3	Septicemia due to anaerobes	6.0×10^{-11}
038.9	Unspecified septicemia	7.2×10^{-11}
238.70	Neoplasm of uncertain behavior of other lymphatic and hematopoietic tissues	3.4×10^{-10}

Phenome Map


2-D Phenome Map, Showing Significant Diagnoses
WBC, Standard Units
0
20
40
60
80
100
ICD-9-CM Code
0
200
400
600
800
1000
1200
1400
cdif

Clostridium difficile

- Bacterial infection, usually brought on through the use of antibiotics
- Symptoms range from mild diarrhea to extreme dehydration, inflammation of the colon, kidney failure, etc
- Infection can be tested for through growing microbiology cultures, but take up to 24 hours to get results. Lab tests results could come back within 5 hours.
- 723 total patients with ICD9 code for *C. diff* in MimicII database
- Phenome map shows *C. diff* to be a highly probable occurence in patients with WBC in the range between 15,000/μL and 45,000/μL

EARLY AND LATE C. *diff*

Clostridium Difficile can occur in two different ways in the ICU:

- Early C. *diff*: Clostridium Difficile infection is found in the patient before admission to the ICU.
- Defined as:
  - Any patients who had a positive microbiology test within 72 hours of admission
  - Patients given treatment within 48 hours of admission. ICD9 code required if treatment was Metrodiazole but no code required if Vancomycin
- Late C. *diff* (Hospital Acquired): Patient acquires C. *diff* infection during ICU stay
  - Positive microbiology test more than 72 hours after admission
  - Order for Vancomycin more than 48 hours after admission
  - Order for Metrodiazole more than 48 hours after admission along with positive ICD9 code

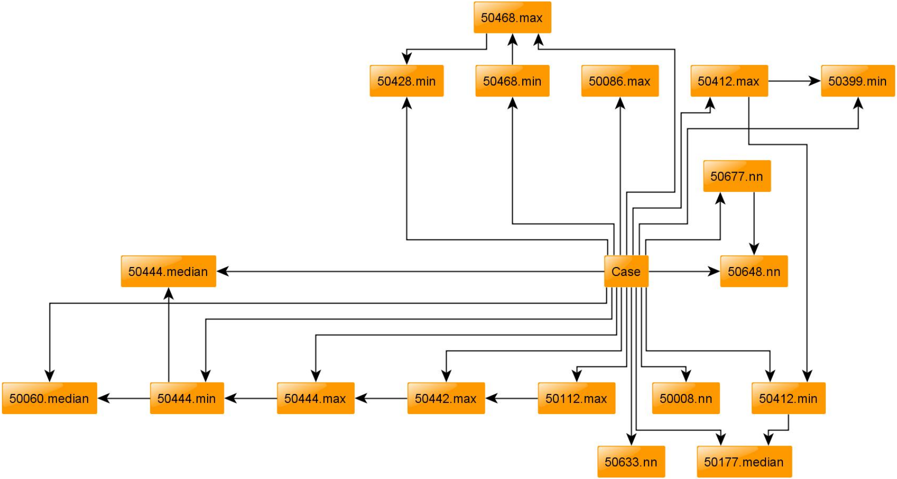
What are Bayesian Networks?

- A Bayesian Network is a graphical model for determining probabilistic relationships among a set of variables. Represented by a directed acyclic graph
- Utilizes machine learning and bayesian probabilities to identify relationships among independent and dependent variables
- Each node is represented by a probability function which determines the probability of a variable given the values of its parents

METHODS

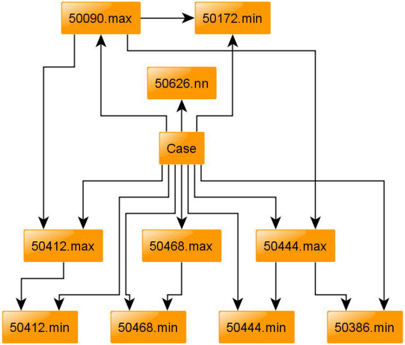
- Bayesian networks generated using WEKA
  - Java based program created by University of Waikato for machine learning
- Patient lab data extracted through taking maximum, minimum, and median lab values throughout duration of hospital stay
- Equal number of negative control patients taken from MimicII
- Data discretized into three equal frequency bins to convert numeric lab data into nominal ranges
- Bayesian network generated through two parent K2 search algorithm and evaluated by 10 fold cross validation
- Attribute selection performed on datasets to reduce the amount of data required for accurate identification

EARLY C. *diff* - ALL LABS

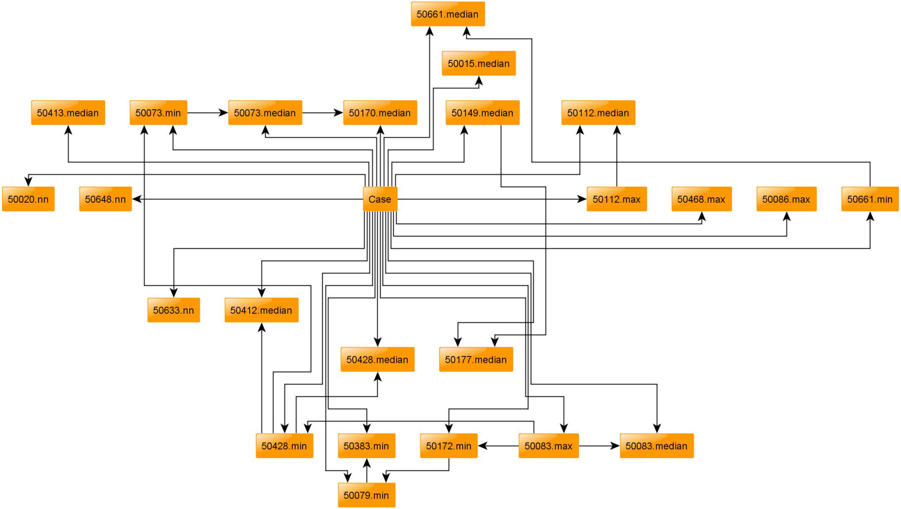


EARLY C. *diff*

```
Full Data
Correctly Classified Instances         593               77.6178 %
Incorrectly Classified Instances       171               22.3822 %

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.801     0.249     0.763     0.801     0.782      0.846    Y
                 0.751     0.199     0.791     0.751     0.77       0.846    N
Weighted Avg.    0.776     0.224     0.777     0.776     0.776      0.846

Attribute Selected (All Labs)

Correctly Classified Instances         551               72.1204 %
Incorrectly Classified Instances       213               27.8796 %

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.751     0.309     0.709     0.751     0.729      0.783    Y
                 0.691     0.249     0.735     0.691     0.713      0.783    N
Weighted Avg.    0.721     0.279     0.722     0.721     0.721      0.783

Attribute Selected (25% NA or less)

Correctly Classified Instances         532               69.6335 %
Incorrectly Classified Instances       232               30.3665 %

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.709     0.317     0.691     0.709     0.7        0.78     Y
                 0.683     0.291     0.702     0.683     0.692      0.78     N
Weighted Avg.    0.696     0.304     0.696     0.696     0.696      0.78
```

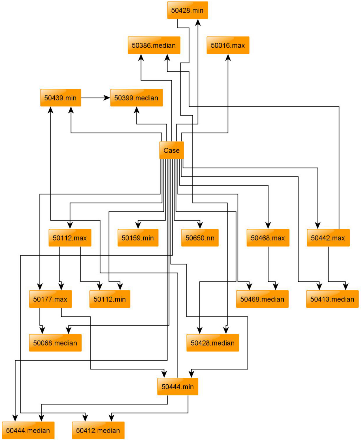


LATE C. *diff* - ALL LAB DATA UP TO TREATMENT

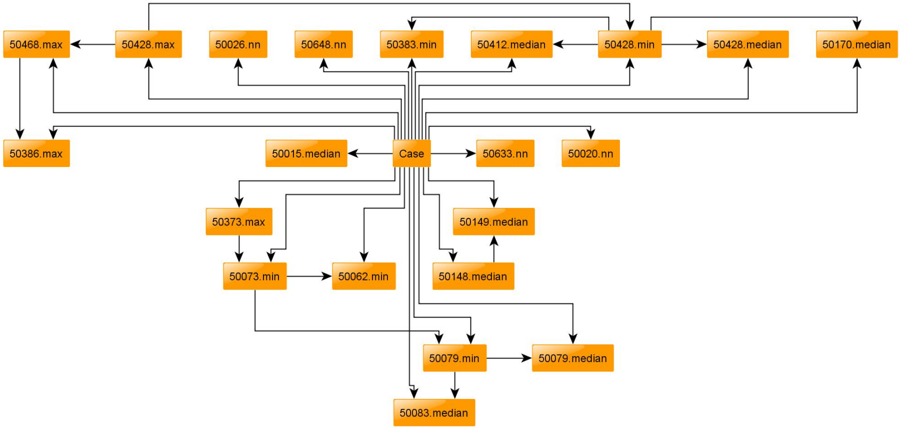


Late C. *diff* - All lab data up to treatment

```
Full Data

Correctly Classified Instances         599               65.9692 %
Incorrectly Classified Instances       309               34.0308 %

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Clas
                   0.584     0.264     0.688     0.584     0.632      0.696    Y
                   0.736     0.416     0.639     0.736     0.684      0.696    N
Weighted Avg.      0.66      0.34      0.663     0.66      0.658      0.696

Attribute Selection (All Labs)

Correctly Classified Instances         547               60.2423 %
Incorrectly Classified Instances       361               39.7577 %

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Clas
                   0.586     0.381     0.606     0.586     0.596      0.656    Y
                   0.619     0.414     0.599     0.619     0.609      0.656    N
Weighted Avg.      0.602     0.398     0.603     0.602     0.602      0.656

Attribute Selection (25% NA or less)

Correctly Classified Instances         528               58.1498 %
Incorrectly Classified Instances       380               41.8502 %

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Clas
                   0.568     0.405     0.584     0.568     0.576      0.621    Y
                   0.595     0.432     0.579     0.595     0.587      0.621    N
Weighted Avg.      0.581     0.419     0.582     0.581     0.581      0.621
```

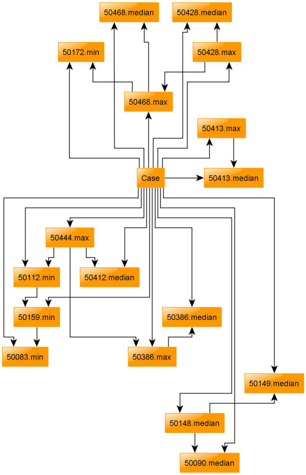


LATE *C. diff* - ALL LAB DATA UP TO 72 HOURS BEFORE TREATMENT



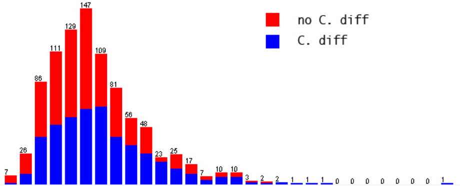
LATE *C. diff* - ALL LAB DATA UP TO 72 HOURS BEFORE TREATMENT

```
Full Data

Correctly Classified Instances         582               66.8966 %
Incorrectly Classified Instances       288               33.1034 %

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.575     0.237      0.708     0.575     0.635      0.722    Y
                 0.763     0.425      0.642     0.763     0.697      0.722    N
Weighted Avg.    0.669     0.331      0.675     0.669     0.666      0.722

Attribute Selected (All Labs)

Correctly Classified Instances         588               67.5862 %
Incorrectly Classified Instances       282               32.4138 %

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.637     0.285      0.691     0.637     0.663      0.739    Y
                 0.715     0.363      0.663     0.715     0.688      0.739    N
Weighted Avg.    0.676     0.324      0.677     0.676     0.675      0.739

Attribute Selected (25% NA or less)

Correctly Classified Instances         563               64.7126 %
Incorrectly Classified Instances       307               35.2874 %

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.584     0.29       0.668     0.584     0.623      0.684    Y
                 0.71      0.416      0.631     0.71      0.668      0.684    N
Weighted Avg.    0.647     0.353      0.65      0.647     0.646      0.684
```



Bayes Net Results

- Results show that our classifiers are able to maintain relatively good levels of accuracy even using minimal amounts of lab tests (≤ 15)
- Accuracies are not perfect in the Bayes nets but are significant enough to provide new information in clinical applications
- Causes:
  - Database has a large percentage of NA, meaning most patients don't get every single lab test taken for them
  - Some errors in database: Impossible values which lead to development of outliers that introduce noise to the data. Much of the lab data has extremely high skew.



CONCLUSION

- Phenome Wide Analysis using WBC allowed us to identify *C. diff* as a commonly occuring diagnosis in a specific range
- Further refining of phenotypic definitions allowed us to identify existance of community acquired and hospital acquired Clostridium difficile
- Bayes net classifiers generated were able to accurately identify Clostridium Difficile cases using minimal lab tests

FUTURE WORK

- Expand method into identifying and finding associations within other phenomes
- Identification of waveforms in lab records
- Inclusion of additional data available in MimicII cohort

ACKNOWLEDGEMENTS

Special thanks to:

- Jeremy Warner and Gil Alterovitz for their many suggestions and advice on the project
- Amin Zollanvari and Swetha Sampath for their assitance with Weka and Bayesian Networks
- PRIMES for providing me this unique research opportunity