

# Natural Language Processing for Biosurveillance

**Wendy W. Chapman, PhD**

Center for Biomedical Informatics

University of Pittsburgh

# Overview

- Motivation for NLP in Biosurveillance
- Evaluation of NLP in Biosurveillance
  - How well does NLP work in this domain?
  - Are NLP applications good enough to use?
- Conclusion

# What is Biosurveillance and Why is NLP Needed?

# Biosurveillance

- Threat of bioterrorist attacks
  - October 2002 Anthrax attacks
- Threat of infectious disease outbreaks
  - Influenza
  - Sudden Acute Respiratory Syndrome
- Early detection of outbreaks can save lives
- Outbreak Detection
  - Electronically monitor data that may indicate outbreak
  - Trigger alarm if actual counts exceed expected counts

# Emergency Department: Frontline of Clinical Medicine

## Triage Nurse/Clerk



### Electronic Admit Data

- Free-text chief complaint
- Coded Admit diagnosis (rare)
- Demographic Information

## Physician



### Electronic Records

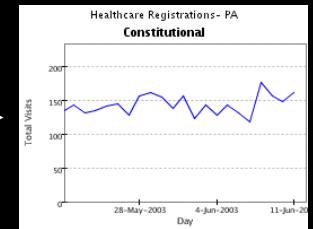
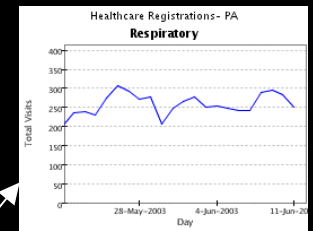
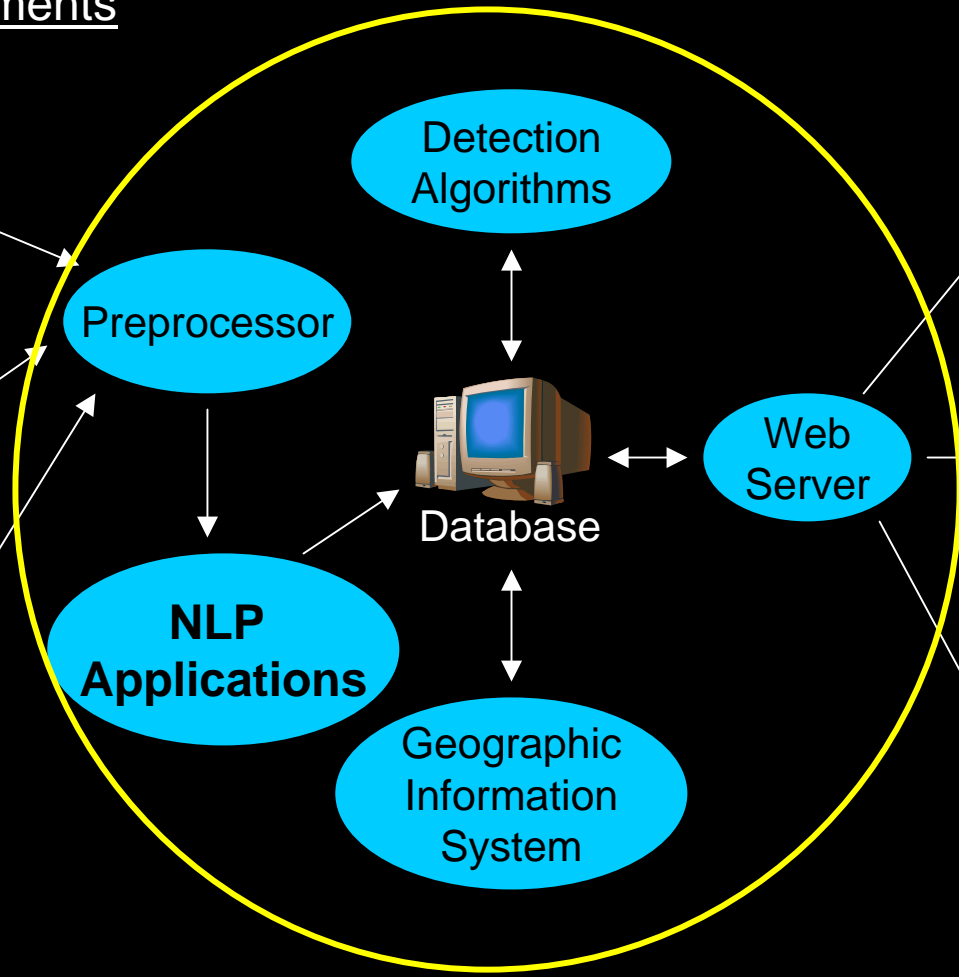
- ED Report
- Radiology Reports
- Laboratory Reports

# RODS System

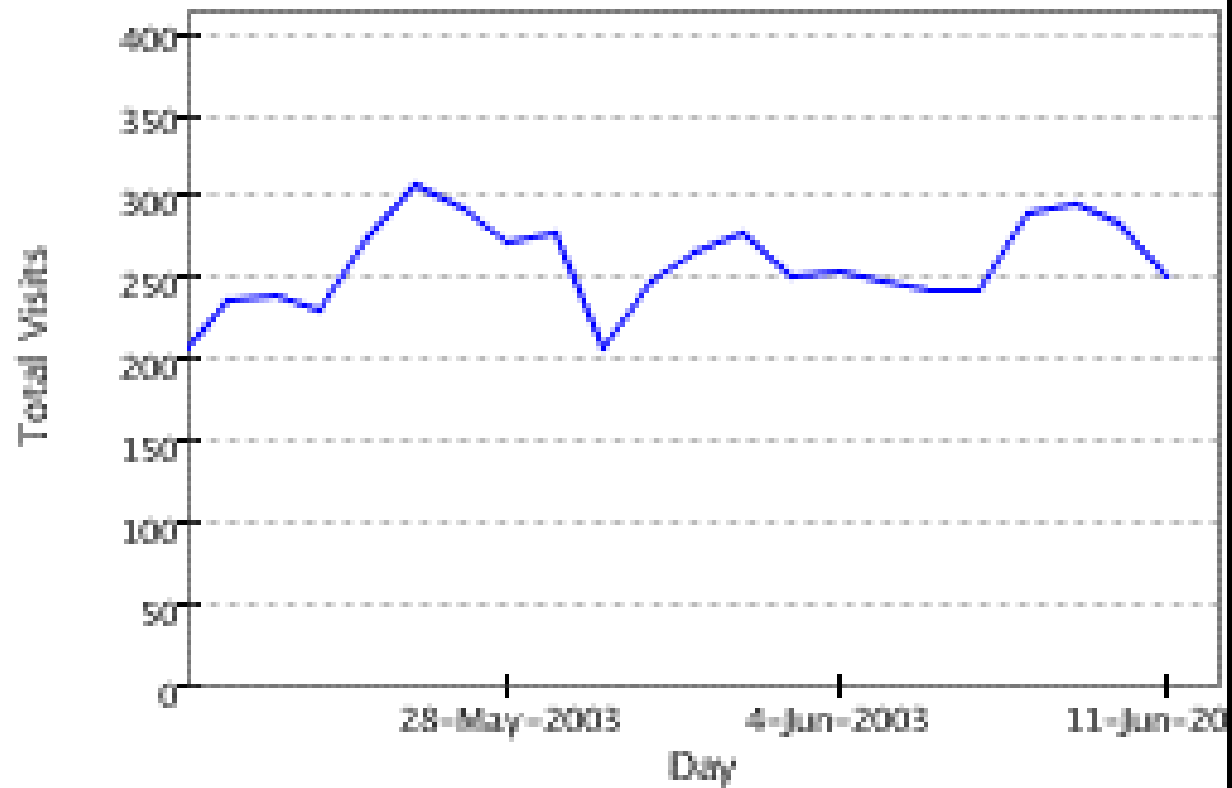
Admission Records from  
Emergency Departments

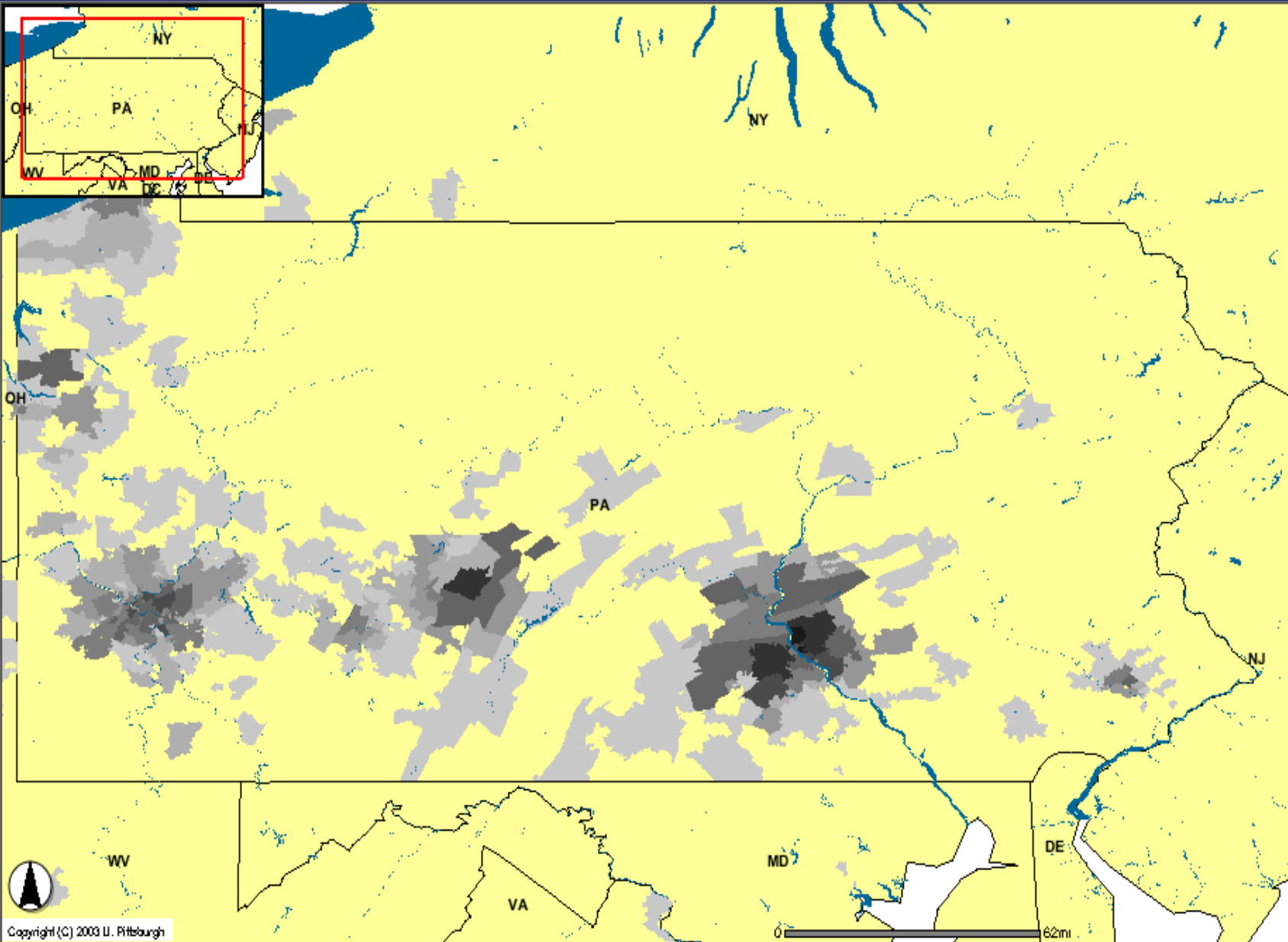
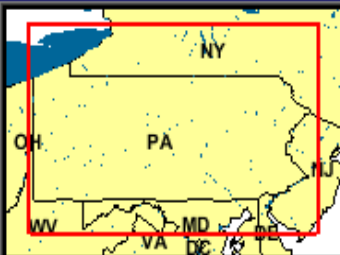
RODS System

Graphs and Maps



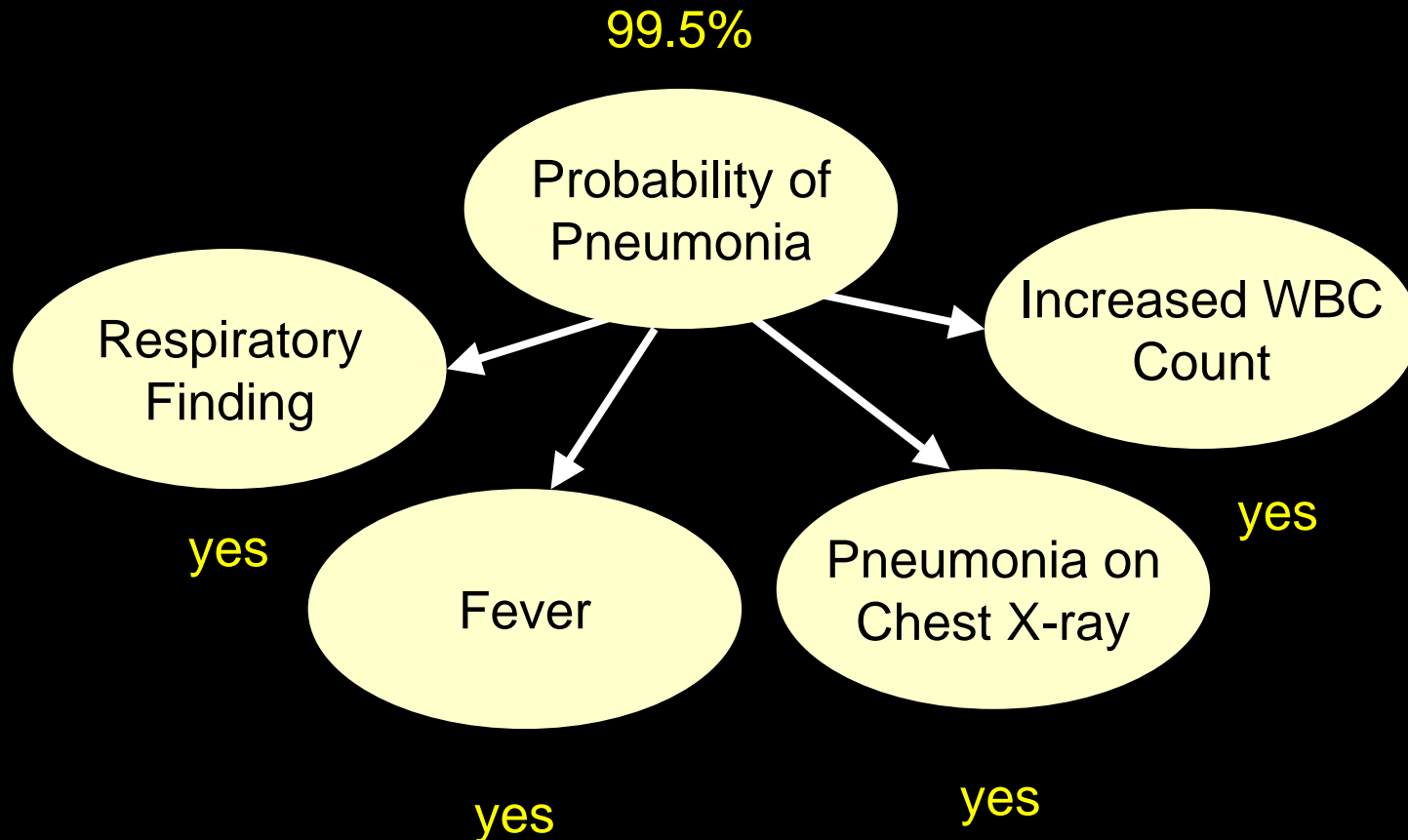
# Healthcare Registrations- PA Respiratory





# Possible Input to RODS

## Pneumonia Cases



# How To Get Values for the Variables

- ED physicians input coded variables for all concerning diseases/syndromes
- NLP application automatically extract values from textual medical records

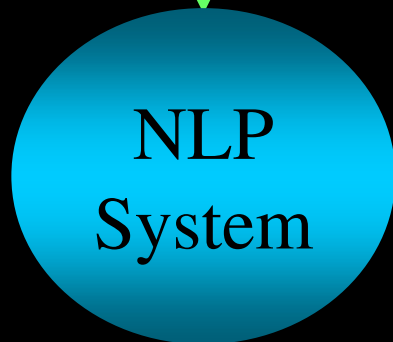
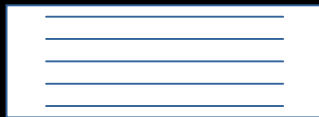
Our research has focused on extracting variables and their values from textual medical records

# Evaluation of NLP in Biosurveillance

# Goals of Evaluation of NLP in Biosurveillance

- **How well does NLP work?**
  - *Technical accuracy*
    - Ability of an NLP application to determine the values of predefined variables from text
  - *Diagnostic accuracy*
    - Ability of an NLP application to diagnose patients
  - *Outcome efficacy*
    - Ability of an NLP application to detect an outbreak
- **Are NLP applications good enough to use?**
  - Feasibility of using NLP for biosurveillance

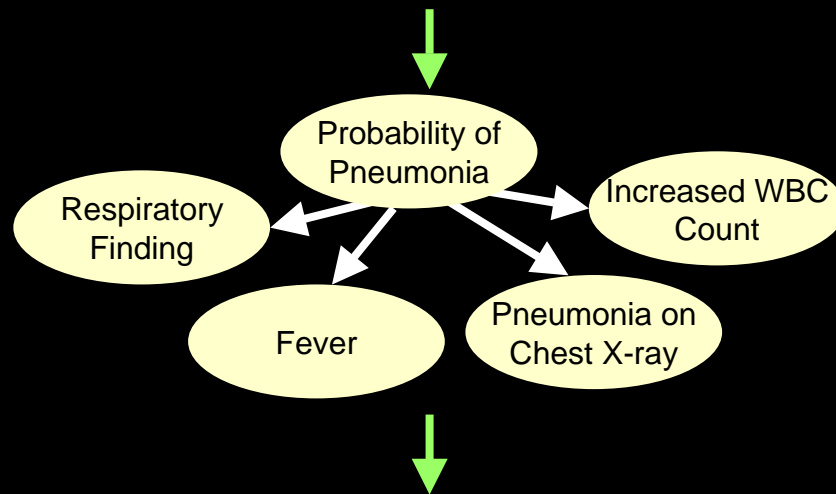
Medical Record



- Respiratory Fx: *yes*
- Fever: *yes*
- Positive CXR: *no*
- Increased WBC: *no*

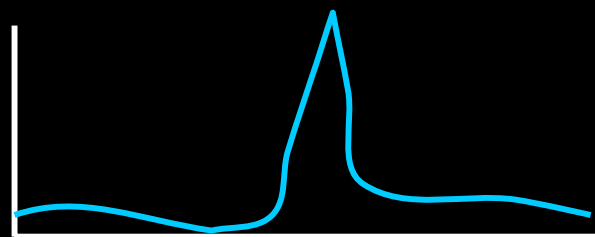
Technical Accuracy

Diagnostic Accuracy



Outcome Efficacy

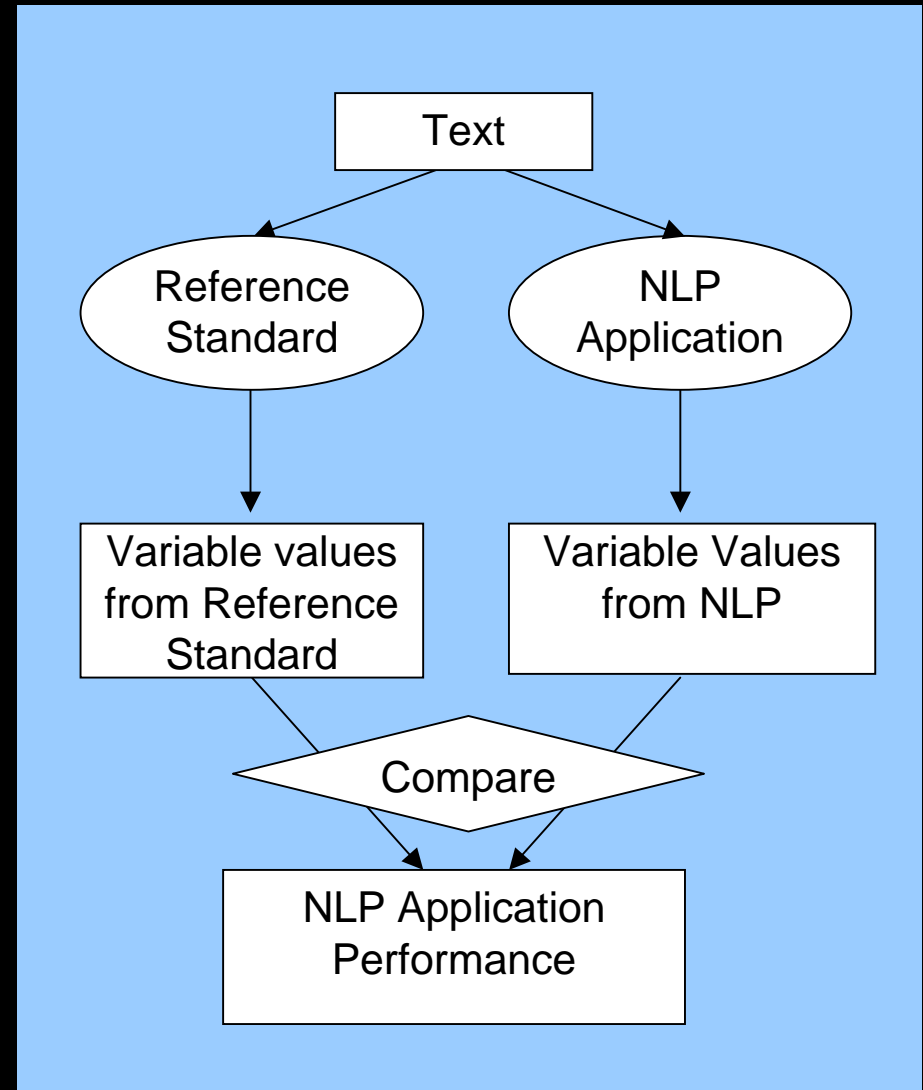
Number of patients with Pneumonia



# Technical Accuracy

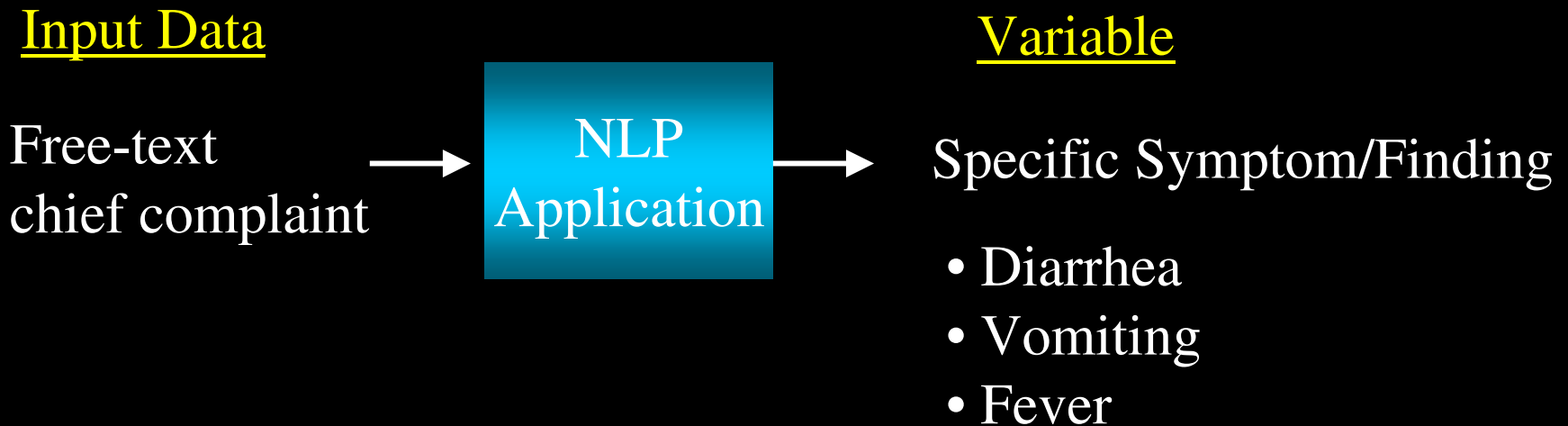
*Can we accurately identify variables from text?*

- **Does** measure NLP application's ability to identify findings, syndromes, and diseases from text
- **Does not** measure whether or not patient really has finding, syndrome, or disease



# Chief Complaints

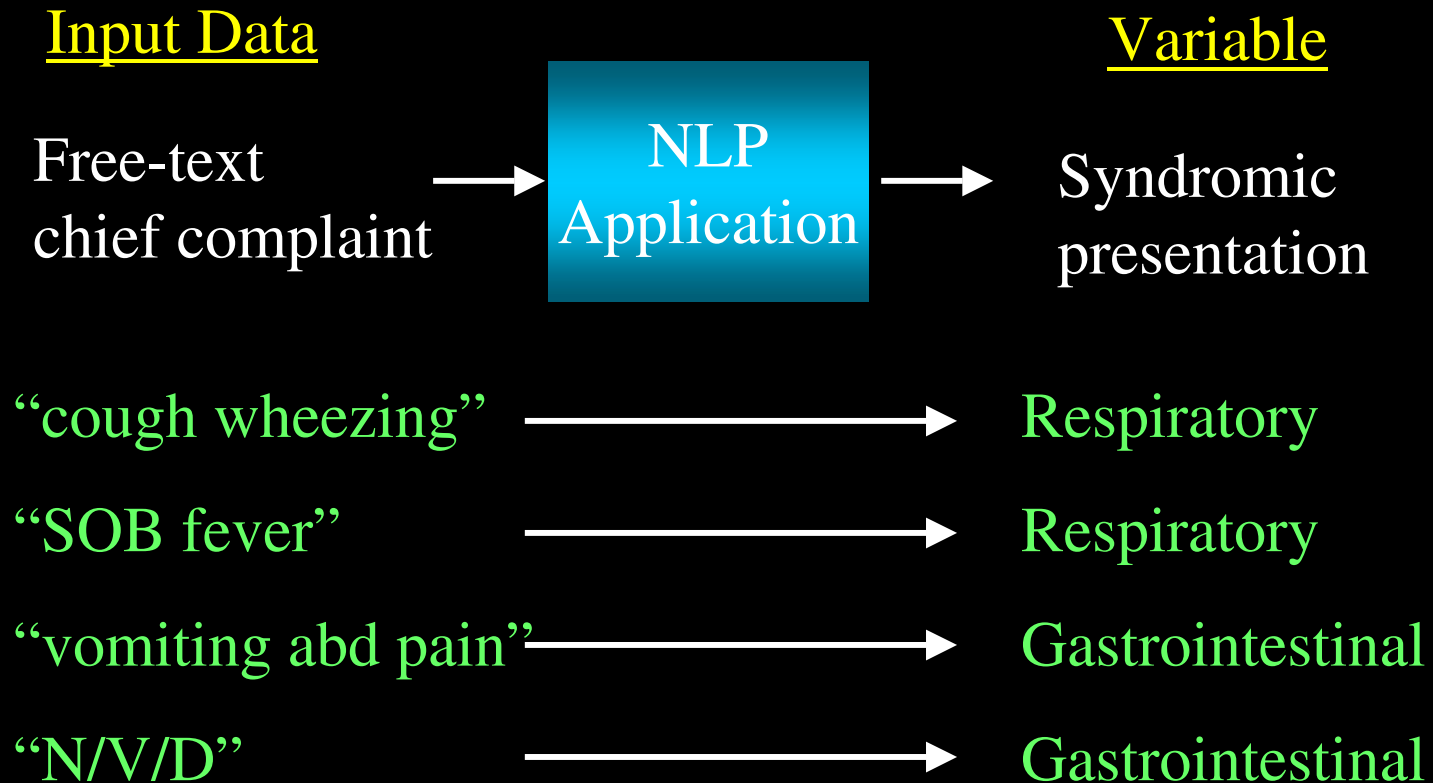
# Extract Findings from Chief Complaints



# Results

	Diarrhea	Vomiting	Fever
Sensitivity	1.0	1.0	1.0
Specificity	1.0	1.0	1.0
PPV	1.0	1.0	1.0
NPV	1.0	1.0	1.0

# Classify Chief Complaints into General Syndromic Categories



# Chief Complaints to Syndromes

## Two Text Processing Syndromic Classifiers

- Naïve Bayesian text classifier (CoCo)\*
- Natural language processor (M+)\*\*

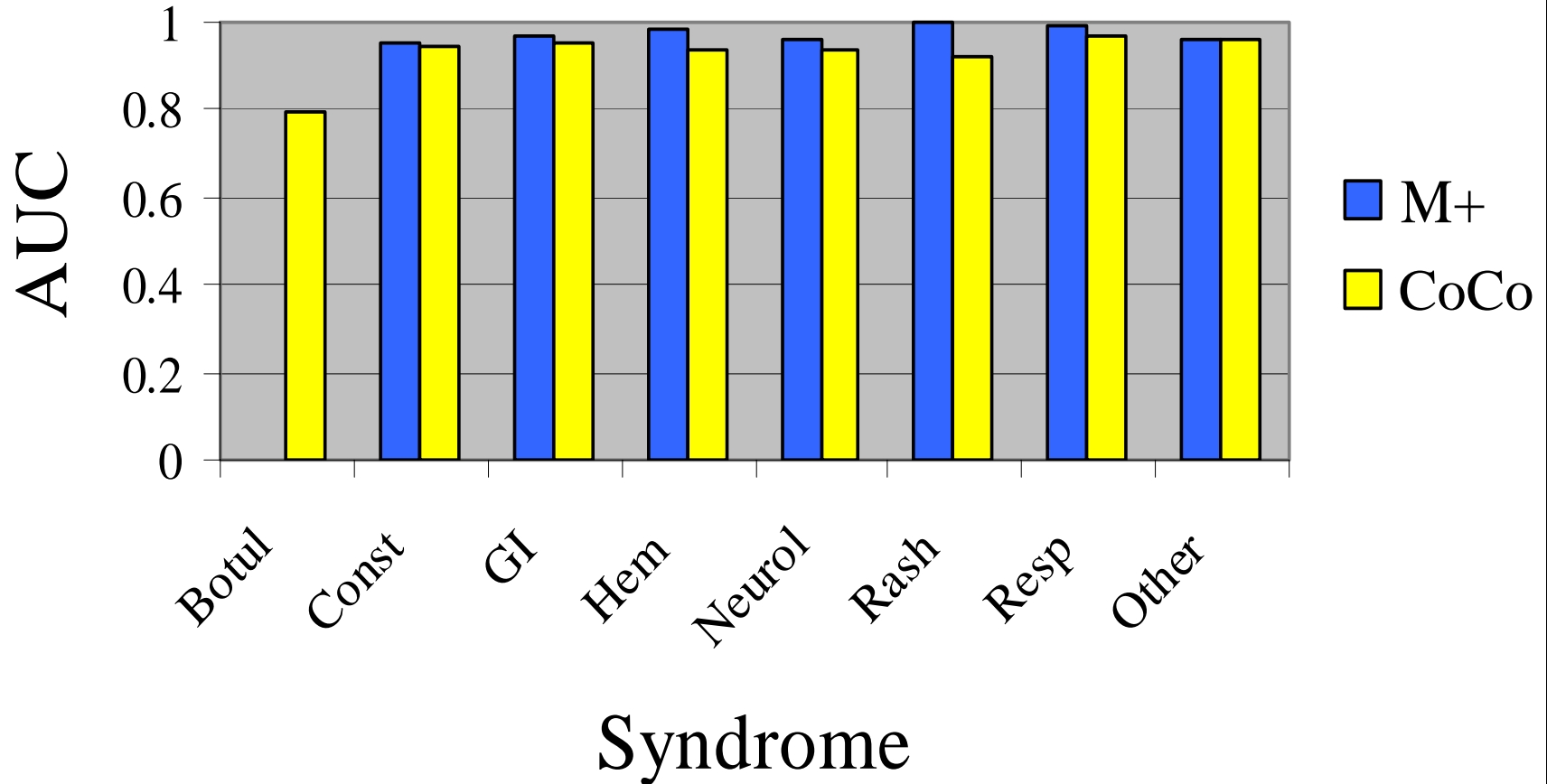
## Methods

- **Task:** classify chief complaints into one of 8 syndromic representations
- **Gold standard:** physician classifications
- **Outcome measure:** area under the ROC curve (AUC)

\* Olszewski RT. Bayesian classification of triage diagnoses for the early detection of epidemics. In: Recent Advances in Artificial Intelligence: Proceedings of the Sixteenth International FLAIRS Conference;2003:412-416.

\*\* Chapman WW, Christensen L, Wagner MM, Haug PJ, Ivanov O, Dowling JN, et al. Classifying free-text triage chief complaints into syndromic categories with natural language processing. AI in Med 2003;(in press).

# Results: Chief Complaints to Syndromes



\* There were no Botulinic test cases for M+

# Chest Radiograph Reports

# Evidence for Bacterial Pneumonia

Detection of Chest x-ray reports consistent with pneumonia			
	Sym-Text	U-KS	<b>P-KS</b>
Sensitivity	0.95	0.87	<b>0.85</b>
Specificity	0.85	0.70	<b>0.96</b>
PVP	0.78	0.77	<b>0.83</b>
NPV			<b>0.96</b>

# Radiographic Features Consistent with Anthrax

## Input Data

Transcribed  
chest radiograph  
report



## Variable

Whether report  
Describes mediastinal  
findings consistent  
with anthrax

- **Task:** classify unseen chest radiograph reports as describing or not describing anthrax findings
- **Gold standard:** majority vote of 3 physicians
- **Outcome measure:** sensitivity, specificity, PPV, NPV

# Mediastinal Evidence of Anthrax\*

Simple Keyword	IPS Model	Revised IPS Model
Sens: 0.043	Sens: 0.351	Sens: 0.856
Spec: 0.999	Spec: 0.999	Spec: 0.988
PPV: 0.999	PPV: 0.965	PPV: 0.408
NPV: 0.979	NPV: 0.986	NPV: 0.999

\*Chapman WW, Cooper GF, Hanbury P, Chapman BE, Harrison LH, Wagner MM. Creating A Text Classifier to Detect Radiology Reports Describing Mediastinal Findings Associated with Inhalational Anthrax and Other Disorders. J Am Med Inform Assoc 2003;10:494-503.

# Emergency Department Reports

# Respiratory Findings

- 71 findings from physician opinion and experience
  - **Signs/Symptoms** – dyspnea, cough, chest pain
  - **Physical findings** – rales/crackles, chest dullness, fever
  - **Chest radiograph findings** – pneumonia, pleural effusion
  - **Diseases** – pneumonia, asthma
  - **Diseases that explain away respiratory findings** – CHF, anxiety
- Detect findings with MetaMap\* (NLM)
- Test on 15 patient visits to ED (28 reports)
  - Single physician as gold standard

\*Aronson A. R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21.

# Detect Respiratory Findings with MetaMap\*

## MetaMap

Sens: 0.70

PPV: 0.55

## Error Analysis

- Domain lexicon
- MetaMap mistake
- Manual annotation
- Contextual Discrimination

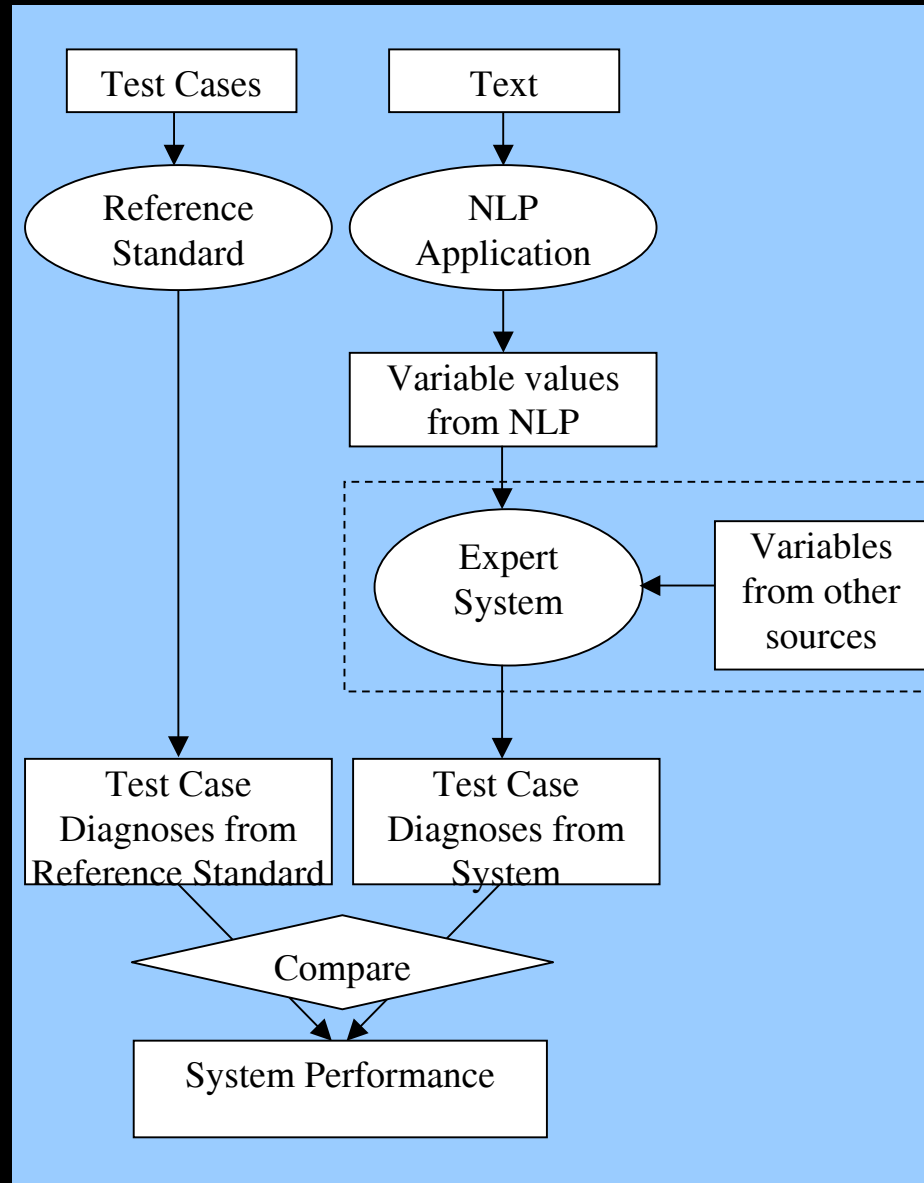
\*Chapman WW, Fiszman M, Dowling JN, Chapman BE, Rindfleisch TC. Identifying respiratory features from Emergency department reports for biosurveillance with MetaMap. Medinfo 2004 (in press).

# Summary: Technical Accuracy

- NLP techniques fairly sensitive and specific at extracting specific information from free-text
  - Chief complaints
    - Extracting individual features
    - Classifying complaints into categories
  - Chest radiograph reports
    - Detecting pneumonia
    - Detecting findings consistent with anthrax
  - ED reports
    - Detecting fever
- More work is needed for generalizable solutions

# Diagnostic Accuracy

*Can we accurately diagnose patients from text?*



# Chief Complaints

# Seven Syndromes from Chief Complaints

- **Gold standard:** ICD-9 primary discharge diagnoses
- **Test cases:** 13 years of ED data

	Positive Cases	Sensitivity	Specificity	PVP
Respiratory	34,916	0.63	0.94	0.44
Gastrointestinal	20,431	0.69	0.96	0.39
Neurological	7,393	0.68	0.93	0.12
Rash	2,232	0.47	0.99	0.22
Botulinic	1,961	0.30	0.99	0.14
Constitutional	10,603	0.46	0.97	0.22
Hemorrhagic	8,033	0.75	0.98	0.43

# Detecting Febrile Illness from Chief Complaints

Technical Accuracy for Fever from Chief Complaints: 100%

## Diagnostic Accuracy

Sensitivity: **0.61** (66/109)

Specificity: **1.0** (104/104)

# Emergency Department Reports

# Detecting Febrile Illness from ED Reports\*

- Keyword search
  - Fever synonyms
  - Temperature + value
- Accounts for negation with NegEx\*\*  
<http://omega.cbmi.upmc.edu/~chapman/NegEx.html>
  - Regular expression algorithm
  - 6-word window from negation term
- Accounts for hypothetical findings
  - *return, should, if, etc.*

**Sensitivity: 98%**

**Specificity: 89%**

\* Chapman WW, Dowling JN, Wagner MM. Fever detection from free-text clinical records for biosurveillance. J Biomed Inform 2004;37(2):120-7.

\*\* Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying Negated findings and diseases in discharge summaries. J Biomed Inform. 2001;34:301-10.

# Summary: Diagnostic Accuracy

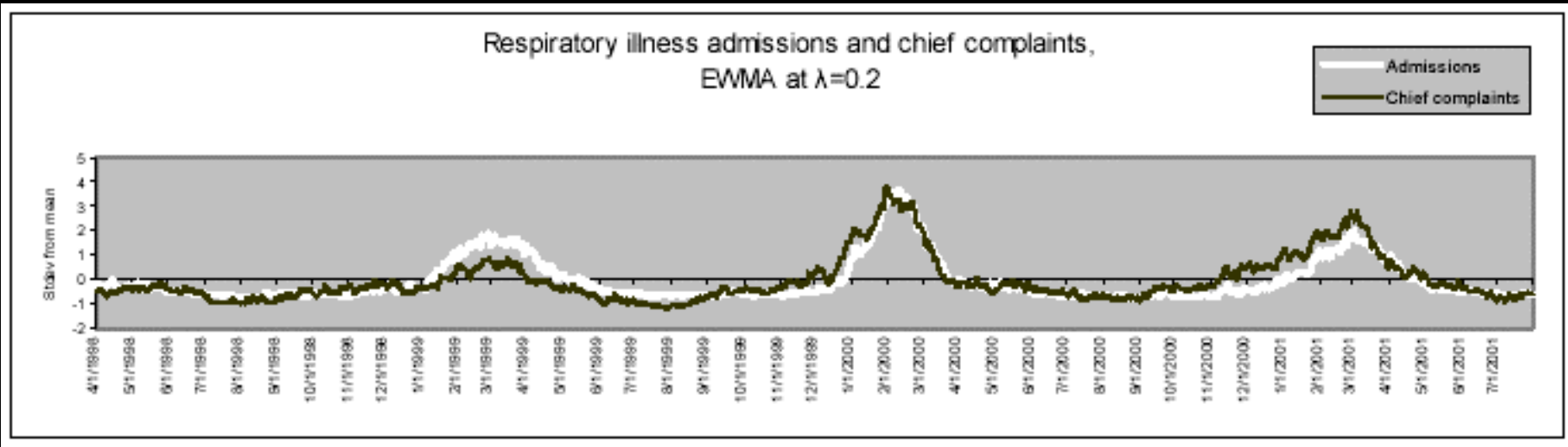
- Good technical accuracy does not ensure good diagnostic accuracy
  - Depends on quality of input data
- The majority of syndromic patients can be detected from chief complaints
- Increased sensitivity requires more information
  - ED reports
- Case detection of one medical problem is doable
  - Fever
- Case detection for more complex syndromes requires more work
  - Pneumonic illness
  - SARS

# Outcome Efficacy

*Can we accurately detect outbreaks from text?*

## Requirements for Evaluation

- Reference standard outbreak
- Textual data for patients involved in outbreak



# Summary: Outcome Efficacy

- Very difficult to test
- Requires trust and cooperation
- Shown that chief complaints contain signal for outbreaks
  - Timelier than ICD-9 codes

# Are NLP Applications Good Enough for Biosurveillance?

## 1. How complex is the text?

- Chief complaints easier than ED reports

## 2. What is the goal of the NLP technique?

- Understand all temporal, anatomic, and diagnostic relations of all clinical findings?
  - Unrealistic
- Extraction of a single variable or understanding of a limited set of variables?
  - Realistic

## 3. Can the detection algorithms handle noise?

- Small outbreaks require more accuracy in variables
  - Inhalational Anthrax outbreak: 1 case = outbreak
- Moderate to large outbreaks can handle noise

# Conclusions

- Patient medical reports contain clinical data potentially relevant for outbreak detection
  - Free-text format
- Linguistic characteristics of patient medical reports must be considered to some extent
- Three types of evaluations necessary to understanding NLP's contribution to biosurveillance
  - How well does NLP works in this domain?
  - How useful are different types of input data?
- Evaluation methods extensible to other domains to which NLP is applied

# Acknowledgments

- Mike Wagner
- John Dowling
- Oleg Ivanov
- Bob Olszewski
- Zhongwei Lu
- Lee Christensen
- Peter Haug
- Greg Cooper
- Paul Hanbury
- Rich Tsui
- Jeremy Espino
- Bill Hogan