

# *Text Normalization for Health Surveillance*

**Alan R. Shapiro, MD**

**Department of Medicine  
New York University School of Medicine**

# Overview

- **Need for text normalization**
- **Two methods of text normalization**
- **Use of text normalization to customize vocabulary and enhance system performance**
- **Performance of text normalization in processing chief complaints**
- **Conclusion**

# Chief Complaint Databases

- **New York City Department of Health and Mental Hygiene Emergency Department Chief Complaint Database**

(DOH) 2,455,598 free text chief complaints

- **Emergency Medical Associates of New Jersey Electronic Medical Record System (EMA)** 3,305,911 free text chief complaints

- **Boston Children's Hospital (AEGIS)** 251,720 free text chief complaints

# Sources of Word Variation in Text Surveillance Data I: Orthographic

<b>SOURCE</b>	<b>EXAMPLES</b>
<b>SPELLING ERRORS</b>	<b>CAUGHING</b>
<b>TYPING ERRORS:</b>	<b>COGHING</b>
Substitutions, Deletions, Insertions	<b>COUGJING</b>
Transpositions, Repetitions	<b>COUHING</b>
<b>GRAMMATICAL VARIATIONS</b>	<b>COUGHED, COUGHS</b>
<b>CONCATENATIONS</b>	<b>COUGHINGALOT</b>
<b>AFFIXES: PREFIXES, SUFFIXES</b>	<b>POSTTRAUMATIC</b>
<b>TRANSCRIPTION ERRORS</b>	<b>CONGHING</b>

# Sources of Word Variation II: Semantic

**LOCAL and IDIOSYNCRATIC USAGE**

**SYNONYMS**

**ACRONYMS, ABBREVIATIONS, TRUNC**

- **Not directly addressed by string matching**
- **~20% of non-stop word strings in chief complaints are abbreviations, acronyms, or truncations.**

# Extent of Variability in Words Often Used in Health Surveillance

NYC DOHMH

<b>Concept</b>	<b># of Representations</b>	<b># of Non-Standard Cases</b>	<b>%</b>
<b>Abscess</b>	<b>92</b>	<b>3419</b>	<b>45.4</b>
<b>Diarrhea</b>	<b>349</b>	<b>4006</b>	<b>11.1</b>
<b>Vomiting</b>	<b>379</b>	<b>14804</b>	<b>16.7</b>
<b>Nausea</b>	<b>137</b>	<b>4143</b>	<b>18.8</b>
<b>Headache</b>	<b>196</b>	<b>1771</b>	<b>3.4</b>

1. Andvomiting	100.Vomitedx5today	300.Vommioting
2. Bomiting	101.Vomiteing	301.Vommitted
3. Cvomiting	102.Vomites	302.Vommiting
---	103.Vomiteted	303.Vommiting
15.V0mitting	104.Vomitfever	304.Vommitintig
16.Vamiting	105.Vomitg	305.Vommitit
17.Vbomiting	---	---
18.Vfomiting	200.Vomitint	325.Vomti
19.Vimit	201.Vomitintg	326.Vomtied
20.Vimited	202.Vomitiny	327.Vomtig
---	---	---
50 Vomiging	250.Vomitting3xdays	377.Vvomitting
51.Vomihing	251.Vomittinga	378.Womiting
52.Vomiig	252.Vomittingab	379.Womitting

# Transcription Errors

<i>r n</i>	<b>UNINATION PREGRANT HENNIA</b>
<i>z g</i>	<b>SEIGURE WHEEGING</b>
<i>n u</i>	<b>CONGHING COUGESTION</b>
<i>l i</i>	<b>PEIVIC MUSCIE LNJURY</b>

# Methods of Text Normalization

**Methods in use focus on string matching:**

- 1. Word stems and keywords- widely used**
- 2. Edit distance methods -number of editing steps needed to transform one string into another.**

# Indexing with word stems has difficulties

**INDEX(CC,"BREAT")+**

**BEATHING, BEEATHING , BRATHING**

**INDEX(CC,"BEATH")+ INDEX(CC,"BRATHING")+**

**INDEX(CC,"DIB")+ INDEX(CC,"D I B")+**

**INDEX(CC,"D.I.B")+INDEX(CC,"DIFF BR")+**

**INDEX(CC,"DIFF, BR")+...**

**BRETHING (32), BRAETHING (19),  
BERATHING (11) and 23 others**

**DIBETES, DIBETIC, DIBETRIC,  
DIBFULATOR and BREATHROUGH**

**Medical logic and word  
normalization need to be kept  
separate**

# Maintenance of Word Stem Systems: Some Disturbing Observations

- **New “words” keep arriving - >750 per week at NYCDOH**
- **Each system has many new word strings to process:  
55% of the “words” in the EMA database, and  
35% of the “words” in the AEGIS database  
are not in the NYCDOH database.**
- **Words used in health surveillance are not spared:  
Diarrhea: 349 ways in the NYCDOH database  
176 additional ways in EMA database.**

# Edit Distance

The minimum cost of transforming one string into another using a defined set of operations (such as insertions, deletions, substitutions and transpositions).

**Asthma** -> **Azthma** -> **Azhma** -> **Azma**  
**Target**                      **Substitution**    **Deletion**        **Deletion**

Problems: “azma” and “stomac”  
score equally close to asthma.

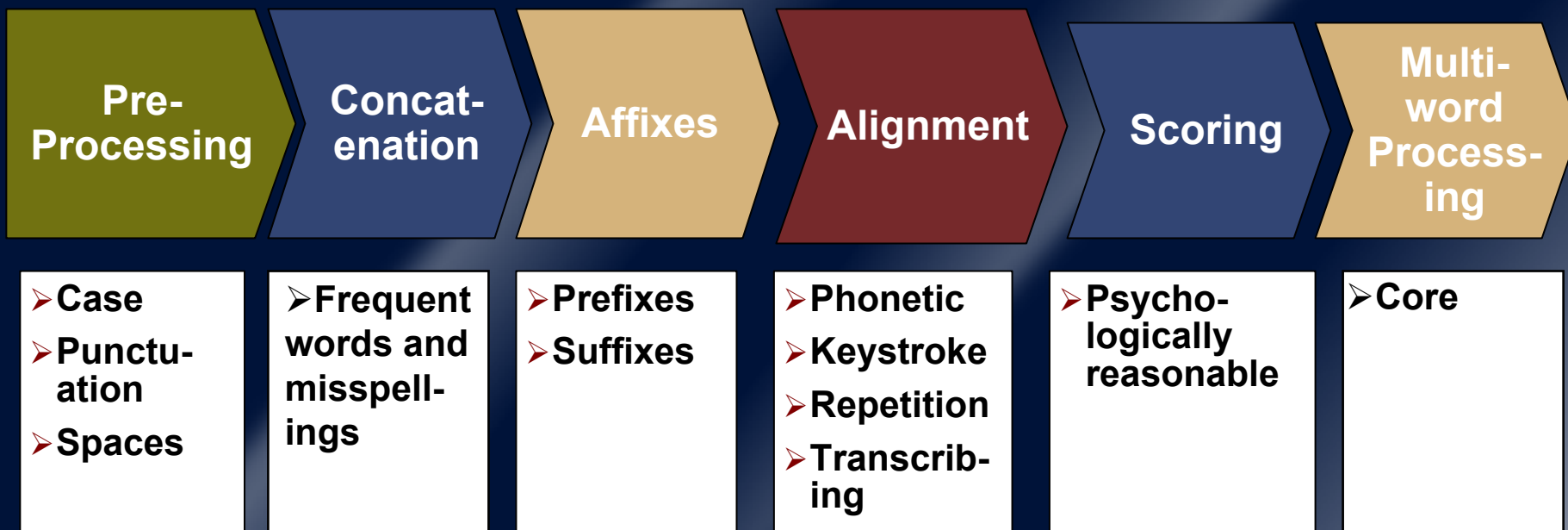
# TextNormalizer

- **TN models the similarity of two strings as the result of a combination of processes:**
  - **Cognitive and performance errors**
  - **Morphological variations**
  - **Affixing medically-related suffixes and prefixes**
  - **Concatenations**
  - **Truncations**
  - **Transcription errors**
- **Approximates psychological distance rather than string edit distance.**

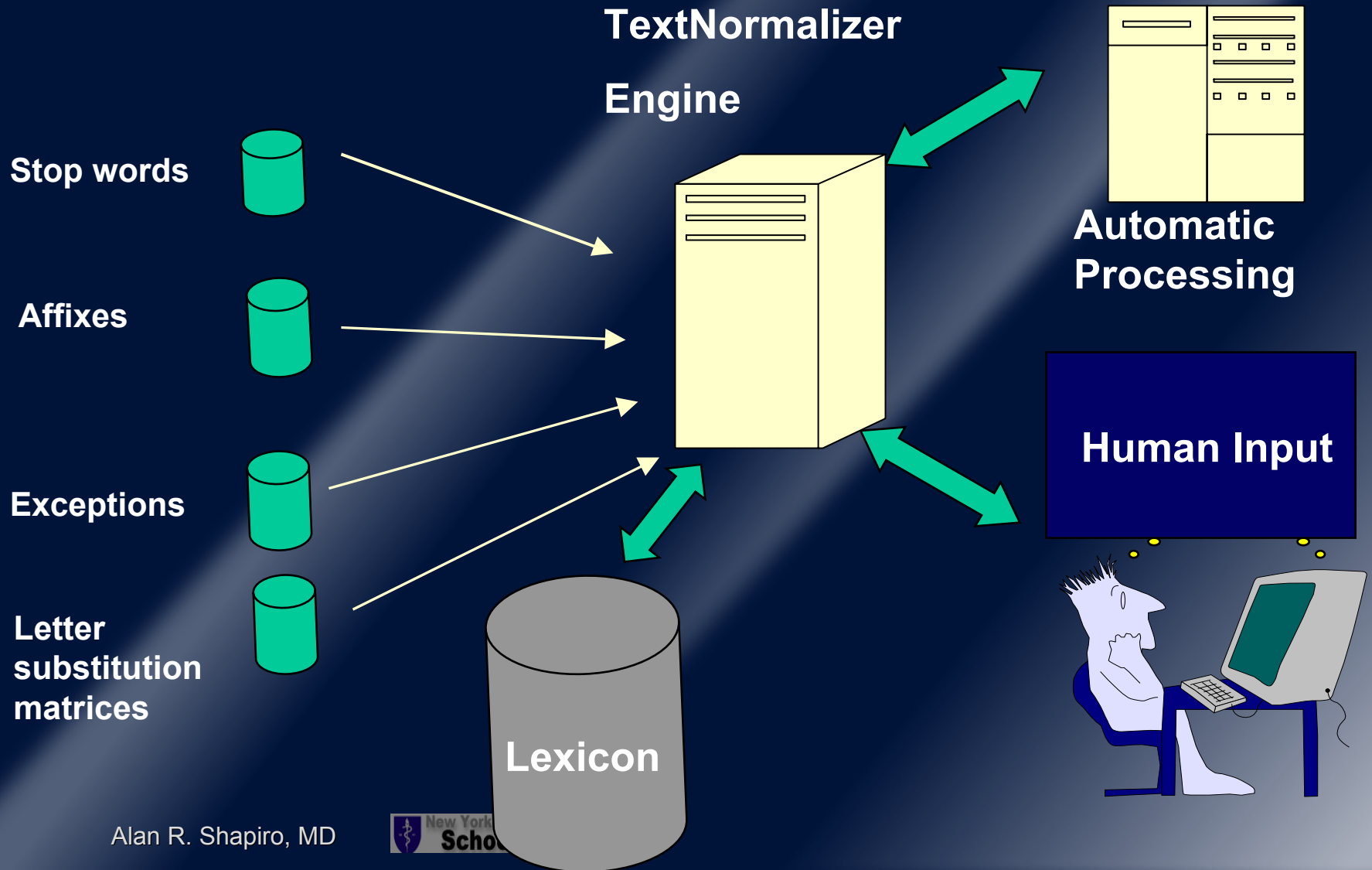
*Is “coughvomitingdiarre” close to “vomiting”?*

<b>Soundex</b>	<b>NO</b>
<b>Edit-distance</b>	<b>NO!</b>
<b>N-gram</b>	<b>SLIGHTLY</b>
<b>TextNormalizer</b>	<b>YES</b>

# The TextNormalizer Engine



# TextNormalizer Architecture



# Use of Text Normalization and ICD-9 Codes to Add Relevant Terms to Syndrome Definitions

Select one or more ICD-9 codes

787.01  
535.50  
535.51  
558.9

Collect the words used in the corresponding chief complaints

DIARR CHLS DIAH DIAHERRIA DYREAH FEVR

\*CRAMPS

\*RUNS

See if there's anything new

Normalize the text

Extract the concepts that are unusually frequent

DIARRHEA  
CHILLS  
FEVER

...

# Candidate words for GI syndrome

<b>Throwing</b>	<b>Poisoning</b>
<b>NonInf</b>	<b>*NVD</b>
<b>*Cramps</b>	<b>*LBM</b>
<b>*Runs</b>	<b>Stomch</b>
<b>*Shigella</b>	<b>Undigested</b>

**\*indicates words identified by normalization that would have been overlooked by current NYCDOHMH algorithms**

# Semantic Normalization

For each word,  
find all the words  
it is associated with



ENTERITIS:

dehydr  
dehydra  
regional  
samonella  
○  
○  
○

REGIONAL

VIRAL

SALMONELLA

ENTERITIS:

BACTERIAL

○  
○  
○

EPIGASTRIC:



Compare every word  
with every other word  
to see which have the  
most similar profiles



Examine the results:



# Performance of Semantic Text Normalization (STN) on some Non-surveillance Words

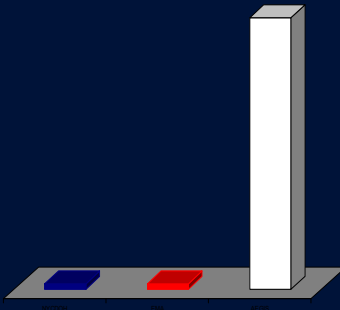
<b>ATRIAL</b>	<b>AFIB</b>
<b>ANKLE</b>	<b>ANKEL, ANKL, ANKELS</b>
<b>CAR</b>	<b>MVC</b>
<b>CUTS</b>	<b>ABRAS</b>
<b>CVA</b>	<b>TIA</b>
<b>DISCHARGE</b>	<b>DISCH</b>
<b>HEROIN</b>	<b>DRUG, HERION, CRACK,DETOX, OPIATE</b>
<b>MOTHER</b>	<b>FATHER</b>
<b>POLICE</b>	<b>NYPD</b>
<b>PUNCH</b>	<b>FIST, BEATEN</b>

# Performance of STN on some Words Used in Health Surveillance

Black	*Dark, brown,*drk,
Cough	*Plegm, *cgh,
Enteritis	*Age,
Fever	Fevr, feve, fev, cough,
Nausea	*N, *NVD, *NV
Pneumonia	RLL, *pneu, exac
Rash	Rashes, *hives,
SOB	DIB
Stool	Urine, dark, brown, black, drk, *tarry

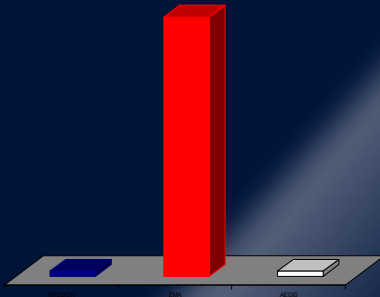
# Local Vocabulary

## AEGIS



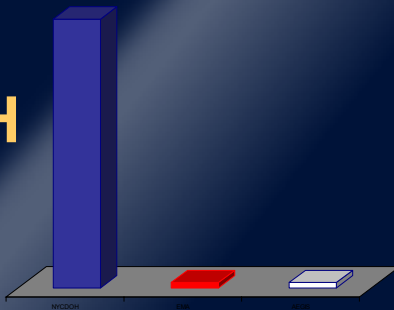
<u>WORD</u>	<u>RR</u>	<u>n</u>
*myalgias	292.2	(102)
*rigors	648.7	(39)

## EMA



*pulm	13.6	(753)
*erupt	66.0	(3392)
*limb	95.5	(12606)

## DOH

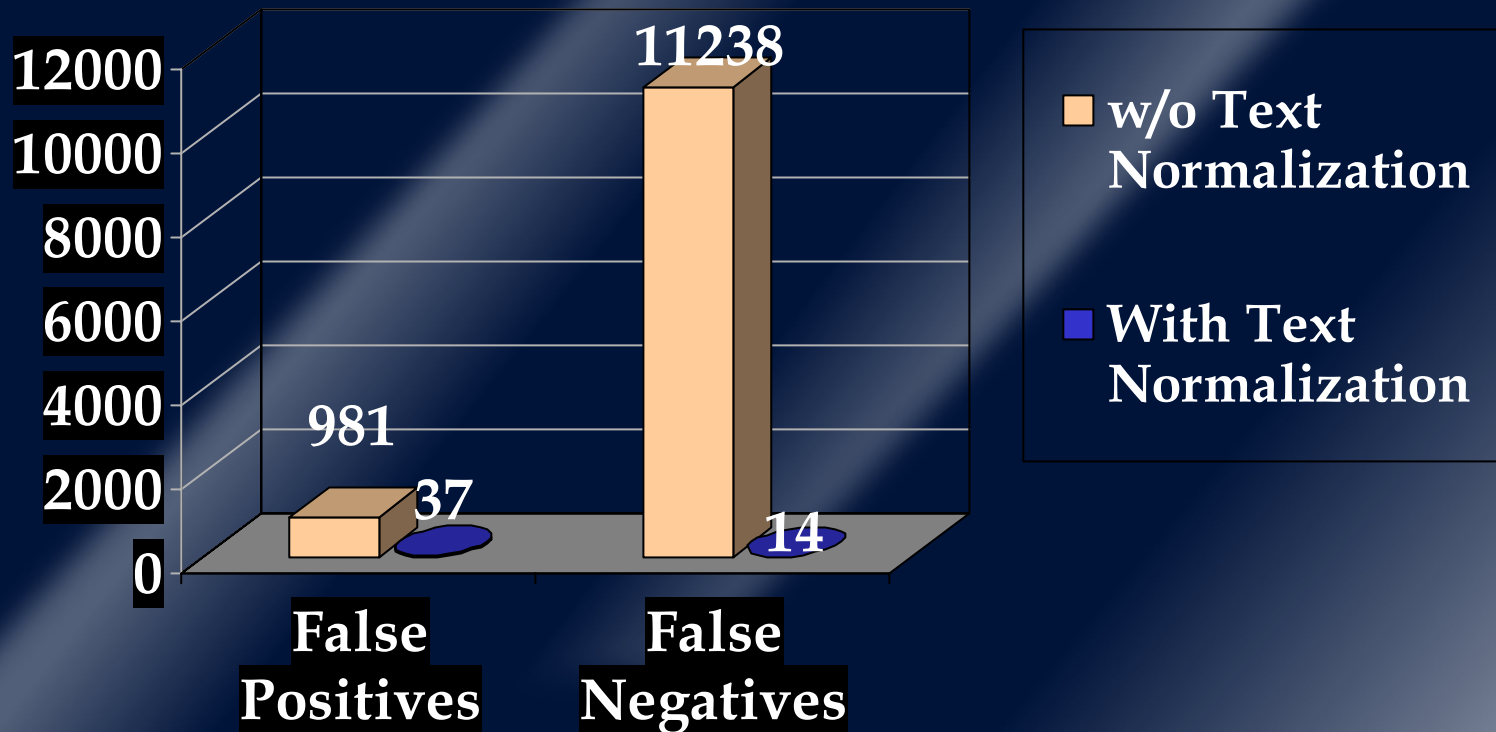


px, pxful	1065.4	(15132)
DIB	1678.3	(2679)
BIB,BIBA	>2000	(11291)

# Error Rates

Syndrome: RESPIRATORY

Database: EMA



**Method:** Word Stems  
**Syndrome:** Respiratory  
**Error:** False Positives (981)

**Word Stem**                      **Retrieves**

---

<b>Breath</b>	“ <b>Breath</b> alizer test ”
<b>DIB</b>	“Swelling L. mandible” ; Drs. <b>Adibe</b> and <b>Dibiase</b>
<b>Monia</b>	“ <b>Ammonia</b> inhalation” ; “ <b>Insumonia</b> -feels tired” ; “ <b>Abdmonial</b> pain”
<b>Resp</b>	“Found <b>unresp</b> ” ; “Needs <b>Respirdal</b> dose”

# Accuracy of estimates of coverage based on extreme values

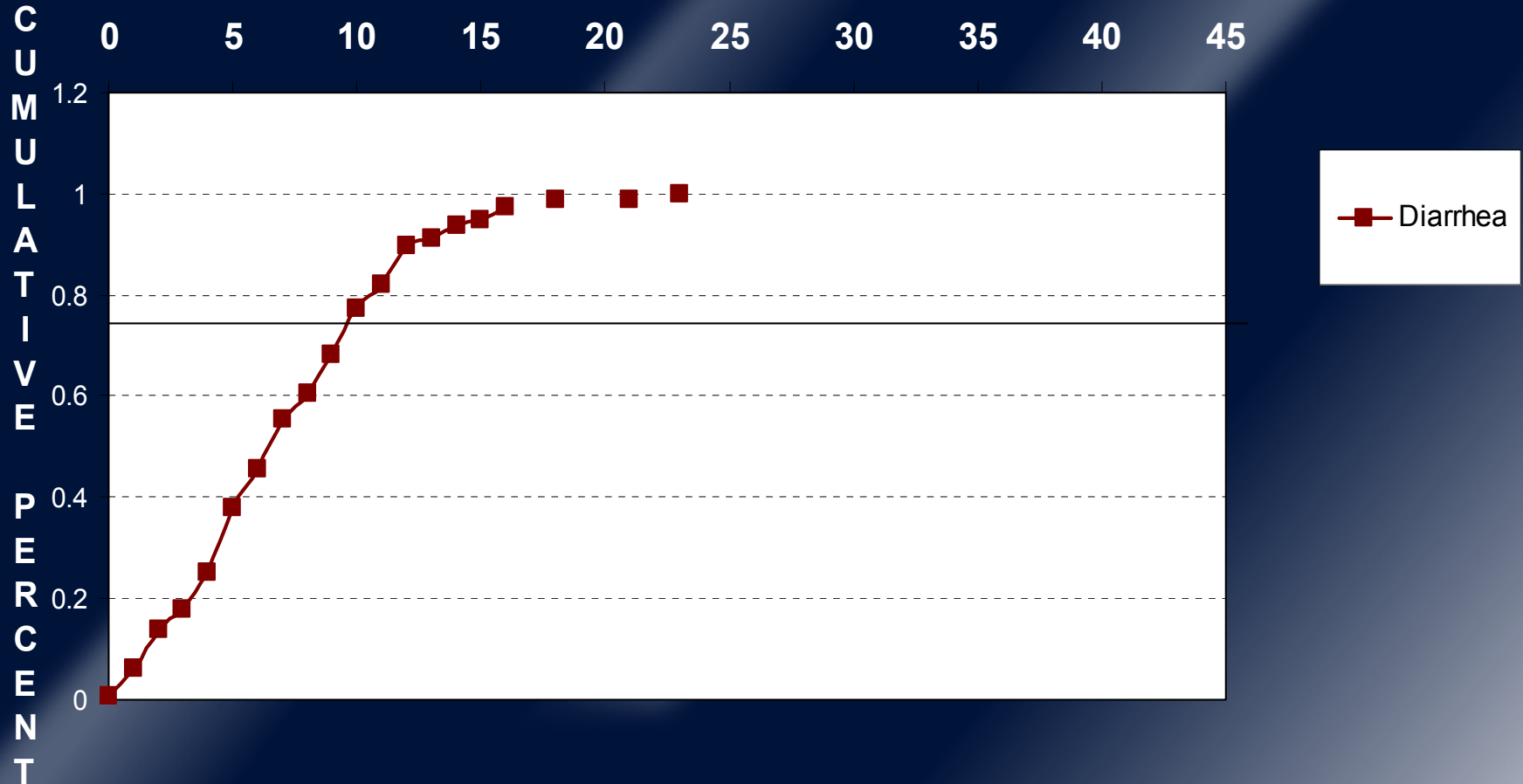
<b>Sample Size</b>	<b>% of times estimate included 99% of population</b>
<b>100</b>	<b>62</b>
<b>1,000</b>	<b>98.4</b>
<b>10,000</b>	<b>99.7</b>

# Small sample comparisons

- 1. For each of 52 weeks, in each of 38 hospitals, calculate the sensitivity in recognizing cases of diarrhea with and without using text normalization.
- 2. For all instances where there were at least 100 cases found, calculate the difference between the two approaches.
- 3. Results: 25% of the time, the difference in sensitivity was 10% or greater.

# Effect of Text Normalization

## INCREASE IN % CORRECT



# Conclusions

- Chief complaints are recorded with an extreme word variability that affects system accuracy.
- Text normalization (TN) is an effective process for:
  - managing word variability
  - identifying relevant vocabulary
  - customizing surveillance systems that use free text