Welcome to the Sentinel Innovation and Methods Seminar Series

The webinar will begin momentarily

- Please visit www.sentinelinitiative.org for recordings of past sessions and details on upcoming webinars.
- Note: closed-captioning for today's webinar will be available on the recording posted at the link above.

Sentinel

Scalable Incident Detection via Natural Language Processing and Probabilistic Language Models

Colin G. Walsh, MD, MA
September 20, 2023

VANDERBILT UNIVERSITY
MEDICAL CENTER

Disclosures

Acknowledgements

Dr. David Kent, Otolaryngology at Vanderbilt University Medical Center (VUMC)

Yufei Long, Research Analyst in my lab

Funding

All investigators were supported on FDA WO2006.

Other disclosures

No relevant conflicts of interest, financial or otherwise, to report

Objectives

- To discuss how Natural Language Processing (NLP) might be used to **detect clinical events** in large-scale health records
- To test whether this approach **generalizes** to diverse events (“phenotypes”)

Background

Table 1. Recent Safety-based Drug Withdrawals

Drug	Adverse Event	Year of Withdrawal
Terfenadine (Seldane)	Drug interactions/arrhythmias (*torsades de pointes*)	1998
Bromfenac (Duract)	Hepatotoxicity	1998
Mibefradil (Posicor)	Drug interactions	1998
Astemizole (Hismanal)	Drug interactions/arrhythmias (*torsades de pointes*)	1999
Grepafloxacin (Raxar)	Arrhythmias	1999
Troglitazone (Rezulin)	Hepatotoxicity	2000
Cisapride* (Propulsid)	Drug interactions	2000
Phenylpropanolamine†	Hemorrhagic stroke	2000
Alosetron (Lotronex)‡	Ischemic colitis	2000
Rapacuronium bromide (Raplon)	Bronchospasm	2001
Cerivastatin (Baycol)	Rhabdomyolysis	2001

* *Cisapride is available only through an investigational limited-access program.*
† *Phenylpropanolamine or PPA was an ingredient in many cough/cold and diet pills.*
‡ *In June 2002, the FDA approved restricted remarketing of alosetron.*
Detailed information on all the above drugs can be accessed by visiting the FDA's website (www.fda.gov/medwatch/safety.htm).

Adverse Drug Event Monitoring at the Food and Drug Administration

Your Report Can Make a Difference

Sved Rizwanuddin Ahmad, MD, MPH

The FDA's safety-surveillance strategy has relied on physicians, health care institutions, manufacturers, and patients to report medical device failures and complications through the Medical Device Reporting system. This system can identify unanticipated medical device failures and complications but requires extensive analytic review and has important limitations.[1] Although the CDRH

Postmarketing Surveillance of Medical Devices — Filling in the Gaps

Frederic S. Resnic, M.D., and Sharon-Lise T. Normand, Ph.D.

N ENGL J MED 366;10 NEJM.ORG MARCH 8, 2012

The FDA's safety-surveillance strategy has relied on physicians, health care institutions, manufacturers, and patients to report medical device failures and complications through the Medical Device Reporting system. This system can identify unanticipated medical device failures and complications but requires extensive analytic review and has important limitations.[1] Although the CDRH

Postmarketing Surveillance of Medical Devices — Filling in the Gaps

Frederic S. Resnic, M.D., and Sharon-Lise T. Normand, Ph.D.

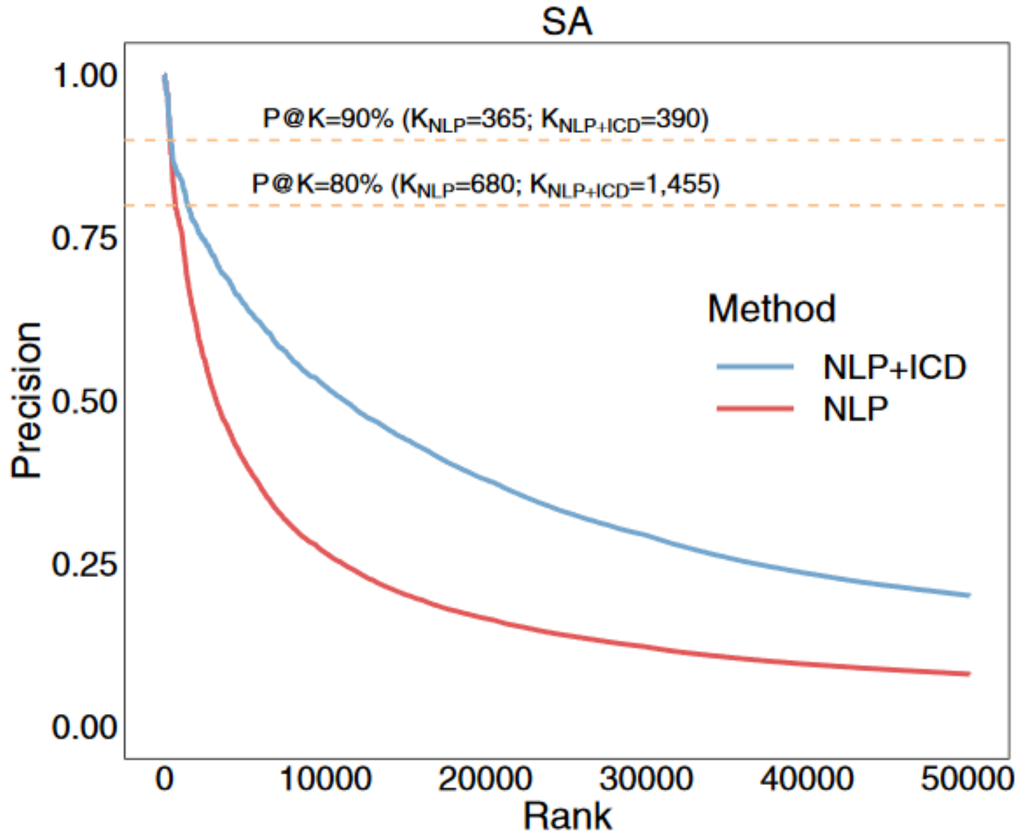
N ENGL J MED 366;10 NEJM.ORG MARCH 8, 2012

Novel systems that:

Detect clinical events at-scale

in a **Timely** fashion and are

Agnostic to complex EHRs



Improving ascertainment of suicidal ideation and suicide attempt with natural language processing

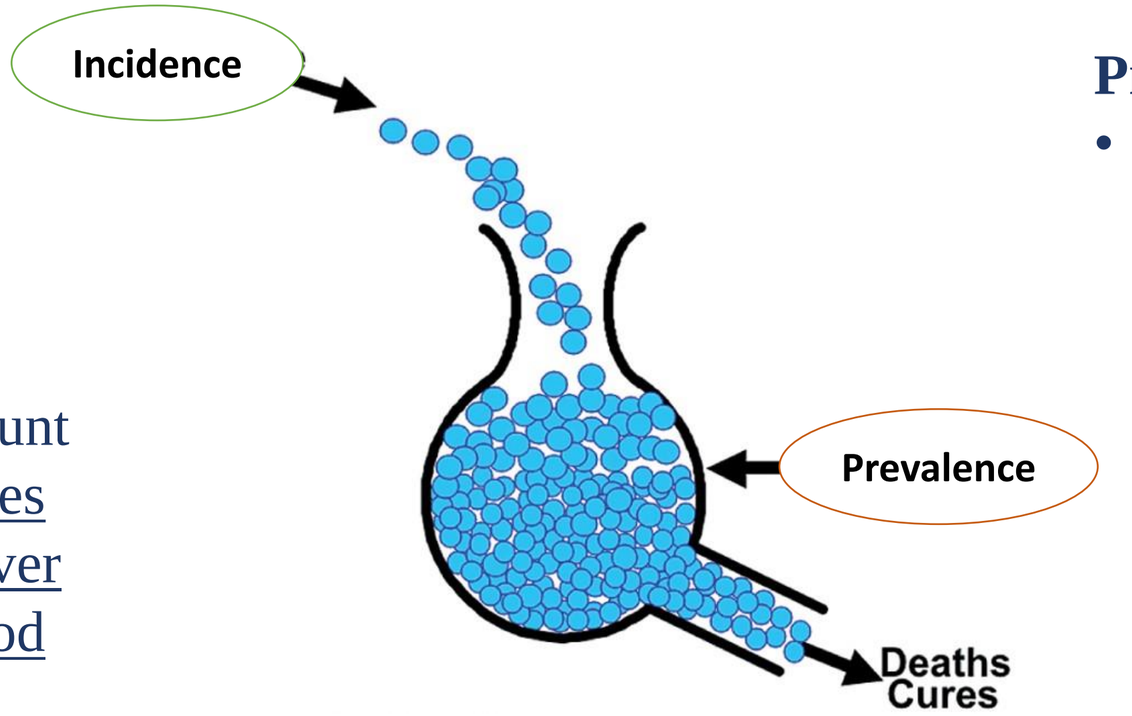
Cosmin A. Bejan[1✉], Michael Ripperger[1], Drew Wilimitis[1], Ryan Ahmed[2], JooEun Kang[3], Katelyn Robinson[1], Theodore J. Morley[3], Douglas M. Ruderfer[1,3,4] & Colin G. Walsh[1,2,4]

Scientific Reports | (2022) 12:15146 | https://doi.org/10.1038/s41598-022-19358-3

Funded by NIMH R01MH121455

Bejan et al measured prevalence of a clinical event

Here, we sought incident clinical events



Incidence

- Describes the amount of new disease cases in at-risk people over a certain time period

Prevalence

- Proportion of population that has a disease or risk factor at a specified point or period in time

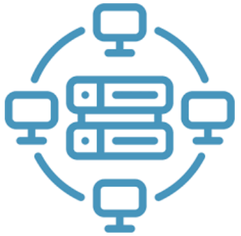
Retrieved from : https://step1.medbullets.com/stats/101004/measures-of-disease-frequency

Other challenges: temporality and lack of interoperability

- Many events have sequelae that result in similar or identical data entry
- Healthcare data might be recorded for a given patient at a later date (e.g., billing) or outside of a “healthcare encounter”

Other challenges: temporality and lack of interoperability

- Many events have sequelae that result in similar or identical data entry
- Healthcare data might be recorded for a given patient at a later date (e.g., billing) or outside of a “healthcare encounter”



- Many open healthcare systems do not have broad interoperability or data sharing
- Care outside one system might not get documented in another

Might NLP improve clinical event detection?

Adapt and test the NLP approach to move from detection across entire records to detecting new, distinct clinical events

Test whether it generalizes (works) in another clinical area

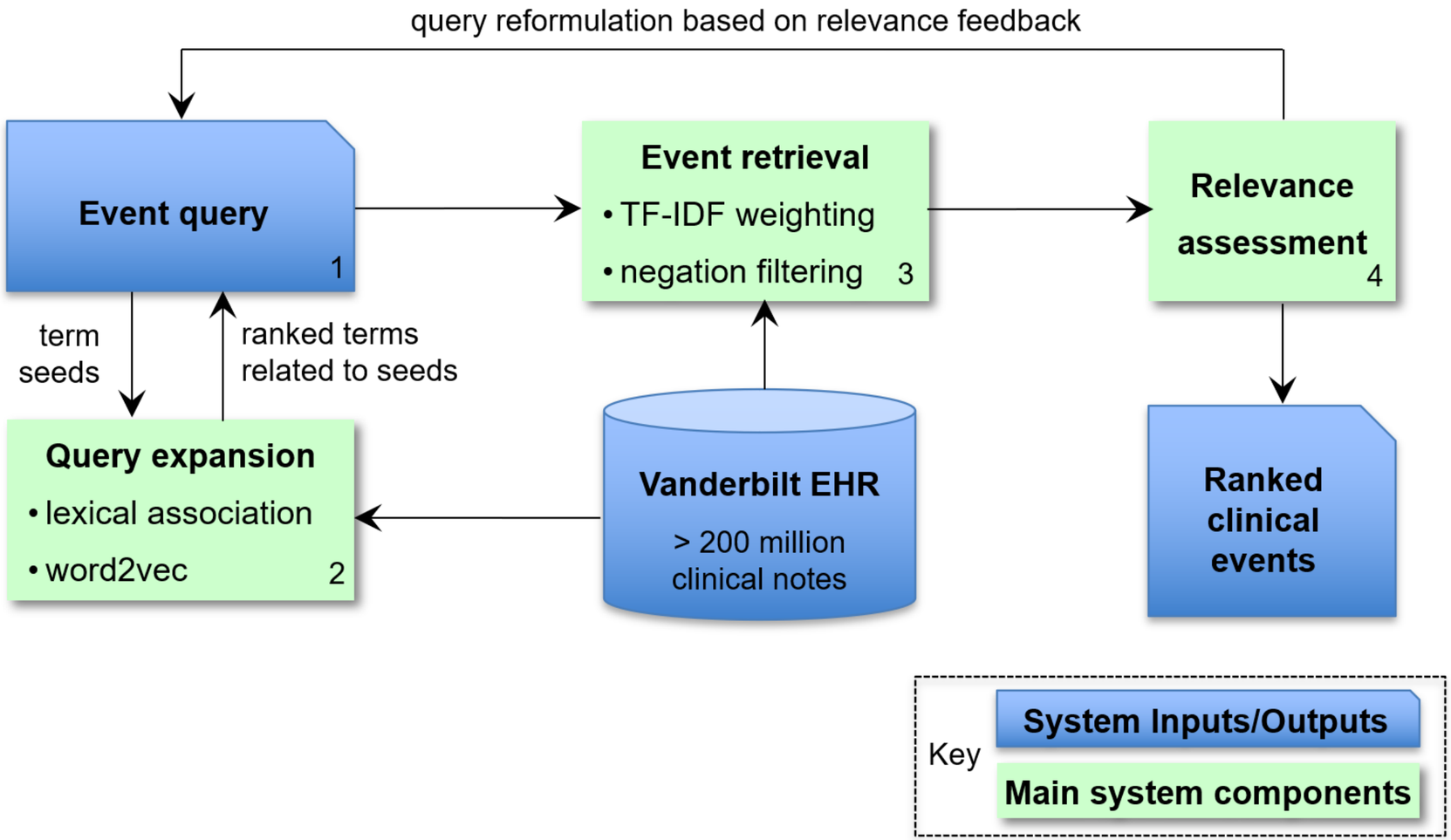
Methods – Study Cohort

- Vanderbilt Research Derivative, research-focused EHR repository, records ranging from 1998 to 2022
- Adult patients aged over 18 years at the time of healthcare encounters with any clinical narrative data in the EHR

Methods –Event Selection

- Suicidal Ideation and Suicide Attempt
- Sleep-related Behaviors

Methods – NLP Approach



Phenotype Retrieval (PheRe) https://github.com/bejanlab/PheRe.git

Methods – Temporality

- Potential temporal windows considered:
  1. Healthcare visit episodes
  2. Set time-windows, e.g., twenty-four hours
  3. Combinations of the above
- We selected a calendar day, e.g., midnight to next midnight, as the window for event detection
  - **Goals: potential utility, simple, and agnostic to EHR**

Methods – Silver Standard

- A source of truth that might be less precise or more error-prone than ground truth, a gold standard
- **Advantage:** relative efficiency to generate and validate compared to more labor-intensive gold standards.

Literature-based diagnostic code sets to measure NLP **preliminary performance**

And to **calculate sample sizes** for chart review (gold standard)

Methods – Gold Standard, Sample Size

- Determine the number of encounters for chart review
  1. Divide predicted risk scores into 5-point intervals and calculating # of encounters per bin
  2. Estimate the precision and recall for each interval using silver standard
  3. Set the marginal error for the probability estimate per bin to 5%

Methods – Gold Standard, Chart Review

- Develop annotation guide for each clinical event
- Conduct training (N~50 chart-days of notes) with multiple reviewers
- Iterate and refine annotation guide
- Conduct multi-reviewer manual chart review noting presence/absence of each phenotype
  - Adjudicate disagreement with third reviewer

Methods – Evaluation Metrics

- Evaluating NLP performance mirrored metrics in preliminary analyses :

1. Precision-Recall (P-R)
2. F1-score

- Calculate error by score bin to understand how well the NLP score performed across all thresholds.

- Replicate a common clinical implementation challenge – discretizing a continuous output from an algorithm into a binary event

Results – Baseline Study Cohort

Characteristic	Suicide Attempt Phenotype (N, %)		Sleep-related Behaviors Phenotype (N, %)	
Total Individuals	89,428		35,863	
Age, median, years	43.2 (95% CI 43, 43.4)		57.7 (95% CI 57.5, 57.9)	
Sex at birth, coded in EHR				
Woman	52,386 (58.6%)		18,850 (52.6%)	
Man	37,003 (41.4%)		17,007 (47.4%)	
Unknown	9 (<1%)		6 (<1%)	
Race, coded in EHR				
White	71113	79.5%	28719	80.1%
Black	11173	12.5%	4341	12.1%
Unknown	4329	4.8%	1880	5.2%
Asian	1756	2.0%	446	1.2%
Other*	901	1.0%	414	1.2%
Alaskan/Native American	156	0.2%	63	0.2%
*Other includes all combinations of coded race categories and a distinct category labeled "Other" in source EHR documentation				
Ethnicity, coded in EHR				
Non-hispanic/Latinx	80,698 (90.2%)		32,832 (91.5%)	
Hispanic/Latinx	2,564 (2.9%)		692 (1.9%)	
Unknown	6,166 (6.9%)		2,339 (6.5%)	

Results

Silver Standard Performance

diagnostic code classification

- Suicide attempt initial benchmarks: 58% PPV for ICD9CM and 85% for ICD10CM

- Sleep-related behaviors: PPV 60% with mix of ICD9/10CM

Results

Silver Standard Performance

diagnostic code classification

- Suicide attempt initial benchmarks: 58% PPV for ICD9CM and 85% for ICD10CM

- Sleep-related behaviors: PPV 60% with mix of ICD9/10CM

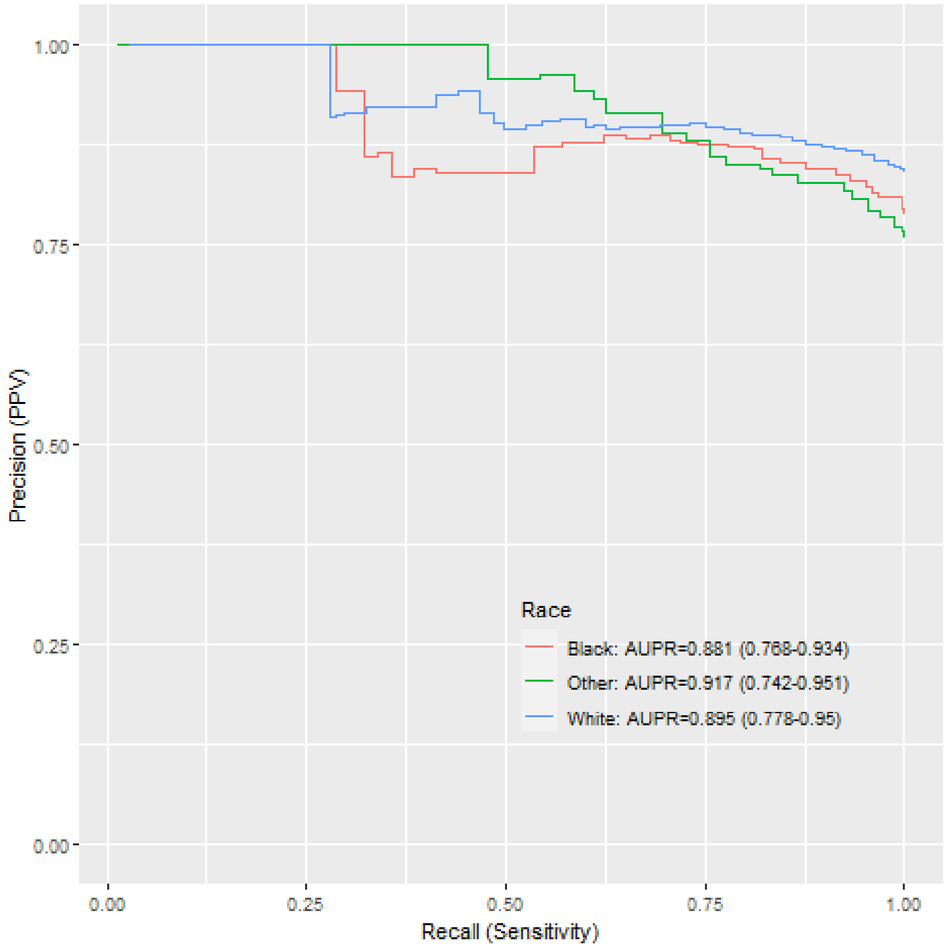
Gold Standard Performance

chart validation

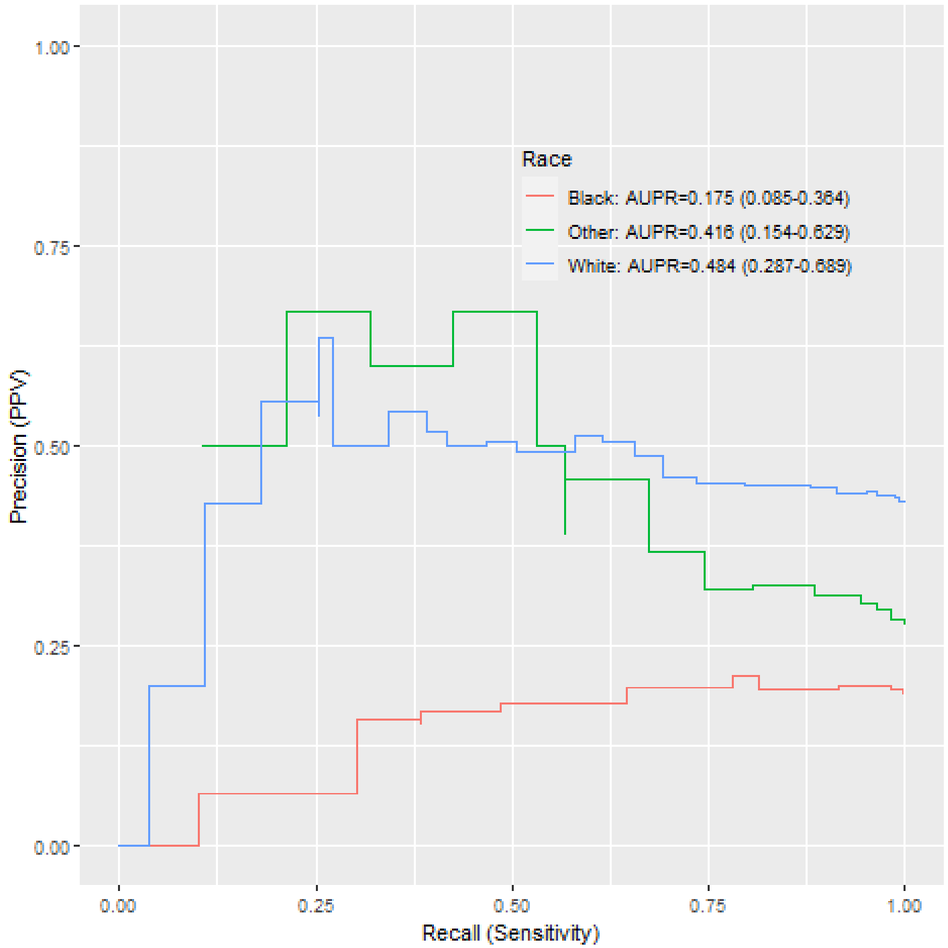
- AUPRC~ 0.77 (95% CI 0.75-0.78) for suicide attempt and AUPRC ~ 0.31 (95% CI 0.28-0.34) for sleep-related behaviors.

Results

Precision-Recall of **Suicide Attempt** NLP Incident Detection by coded race



Precision-Recall of **Sleep-related Behaviors** NLP Incident Detection by coded race



Results

Threshold selection & Steps toward implementation

- Selecting a threshold for a hypothetical implementation where precision matters more than recall, we use the maximum F1-score.

- We note that this score, while the optimal model performance, might not represent an acceptable or optimal performance for specific applications e.g., defining an endpoint of interest in medical product safety surveillance.

- Suicide attempt

NLP score of 25 and higher would have an F1-score of 0.75 associated with a recall of 0.93 and a precision of 0.63.

- Sleep-related behaviors

The optimal F1-score-based threshold is 35 or above in which the F1-score would be 0.42, precision 0.33, recall 0.57.

Discussion

NLP-based clinical event detection was feasible, but performance differed by phenotype.

Even with imperfect coded race variables, performance differences were easily identifiable.

Algorithmovigilance remains paramount in the evolution of systems like this one – not solely at initial algorithm validation but throughout the life cycle.

Invited Commentary | Health Informatics

Algorithmovigilance—Advancing Methods to Analyze and Monitor Artificial Intelligence–Driven Health Care for Effectiveness and Equity

Peter J. Embi, MD, MS

JAMA Network Open. 2021;4(4):e214622. doi:10.1001/jamanetworkopen.2021.4622

Discussion – Performance differences

- Clinical events differed by observability and quality/degree of documentation.
- Differing event rates at baseline.
- Differing rates of diagnostic coding.
- Critical step to determine demographics or other clinical attributes that might undermine successes of an event detection system like this one.

Discussion – Implications

- Silver standard used here for sample size calculation and preliminary performance estimates – **necessary but not sufficient.**
- Human input still indicated in the validation of such systems; not yet fully automated.

Discussion – Future work

- Expanding similar approaches to new phenotypes and replicating pipeline.
- Testing generalizability in new settings and health systems.
- Testing new algorithms (e.g., large language models).
- Considering performance benchmarks prior to implementation.
- Designing key components of over-arching detection system informed by this approach.

Thank you! Questions?

Colin.Walsh@vumc.org

www.walshscience.com

Vanderbilt
Adi Bejan
Cindy Chen
Daniel Fabbri
Kevin Johnson
Jhansi Kolli
Michael Matheny
Michael Ripperger
Katelyn Robinson
Doug Ruderfer
Colin Walsh
Drew Wilimitis
Aileen Wright

KPWA
David Carrell

Harvard Pilgrim
Sruthi Adimadhyam
Elizabeth Messenger Jones

FDA
Sai Dharmarajan
Yong Ma
Andy Mosholder
Danijela Stojanovic

Funding: WO2006

Supplemental Slides

Methods

Automatic extraction of phenotypic profiles from clinical notes

- The method involved processing large collections of clinical notes from EHRs (including tokenization and extraction of n-gram representations such as unigrams and bigrams)
- Unsupervised training of Google's word2vec and Transformer-based NLP models such as Bidirectional Encoder Representations from Transformers (BERT)
- The extraction of phenotypic profiles- iteratively expanding an initial set of high-relevant expressions (seeds)
- Rank the learned embeddings by their similarity to the seed embeddings
- Manually review the top ranked expressions and selected the relevant ones as new seed expressions

Methods

Large-scale retrieval of incident phenotypes

- We implemented a search engine to identify incident phenotypes in all the notes from the Vanderbilt Research Derivative and to rank them by relevance to their profile. In this context, each phenotypic profile corresponds to an input query for the search engine while each meta-document comprising of all the notes of a patient on a given day encodes a potential incident phenotype.
- In the implementation framework, we represented the meta-documents and input queries as multidimensional vectors, where each vector element is associated with a single- or multi-word expression from their corresponding phenotypic profile. The relevance of a patient meta-document to a phenotype was measured as the similarity between their meta-document and input query vectors using the standard term frequency-inverse document frequency (TF-IDF) weighted cosine metric. The final NLP-based score was a continuous value ranging from low single digits (< 10) to hundreds (< 500 typically). Higher scores indicated more similarity and therefore more evidence of the phenotype.
- To further improve the performance of our search engine, we performed query reformulation based on relevance feedback by iterative assessment of the top 20 retrieved incident phenotypes of each run. The selection and ranking of the incident phenotypes was performed using the Phenotype Retrieval (PheRe) software package in Java, which is available at https://github.com/bejanlab/PheRe.git.

Workgroup Members

Vanderbilt
Adi Bejan
Cindy Chen
Daniel Fabbri
Michael Matheny
Michael Ripperger
Katelyn Robinson
Colin Walsh (WG Lead)
Drew Wilimitis
Kevin Johnson
Aileen Wright
Jhansi Kolli
Doug Ruderfer

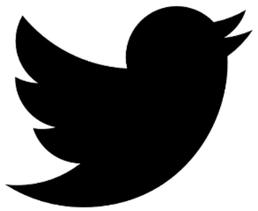
KPWA
David Carrell

Harvard Pilgrim
Sruthi Adimadhyam
Elizabeth Messenger Jones

FDA
Sai Dharmarajan
Andy Mosholder
Danijela Stojanovic
Yong Ma

Thank you! Questions?

Colin.Walsh@vumc.org



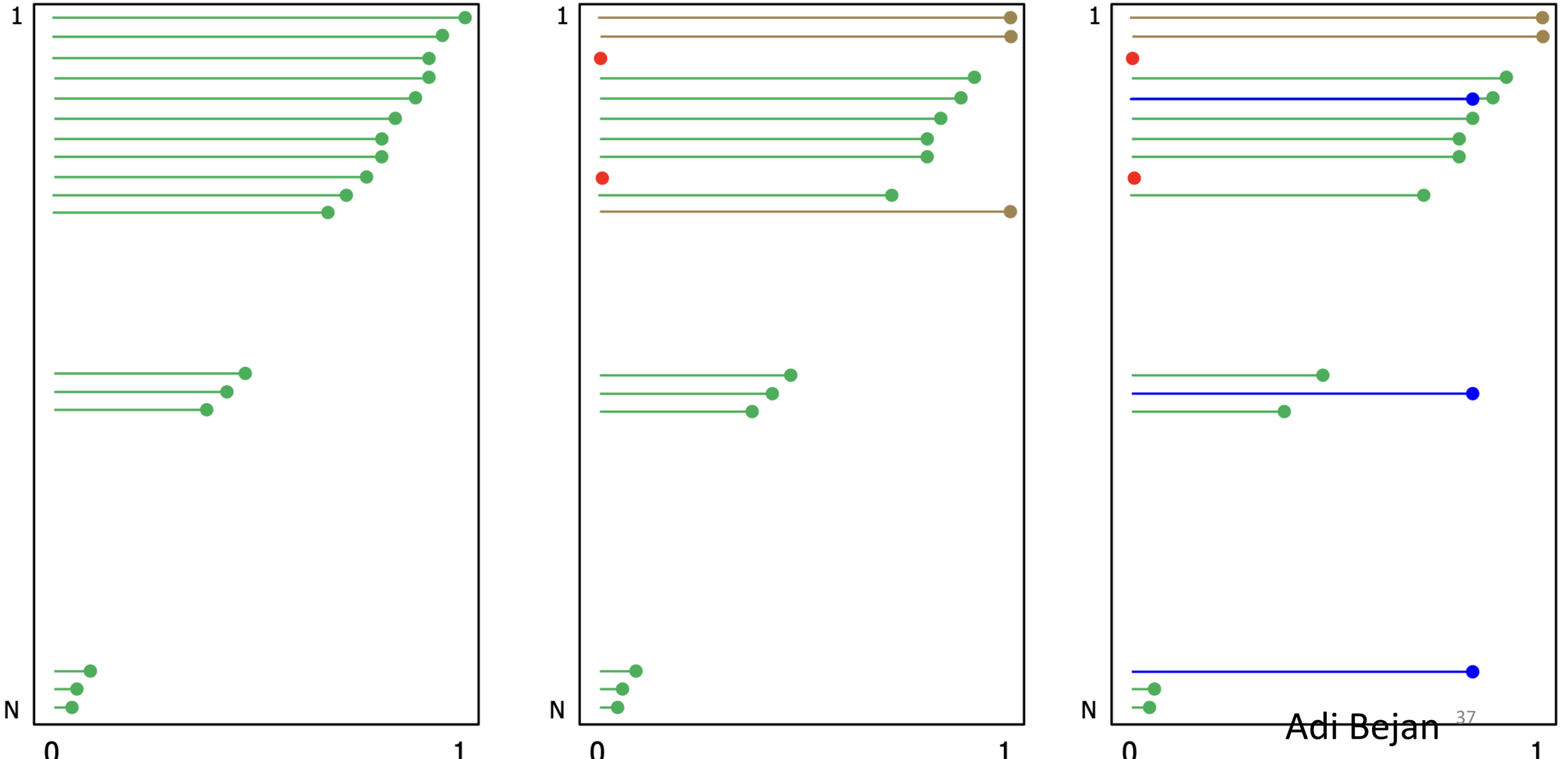
@CWalshMD

www.walshscience.com

P(NLP)

P(NLP)
+ gold labels

P(NLP+ICD)
+ gold labels



Adi Bejan