Improving the Efficiency of Outcome Validation in the Sentinel System

Defining the Problem

Robert Ball, MD, MPH, ScM
Deputy Director
Office of Surveillance and Epidemiology
Center of Drug Evaluation and Research

Case Classification the Old-fashioned Way

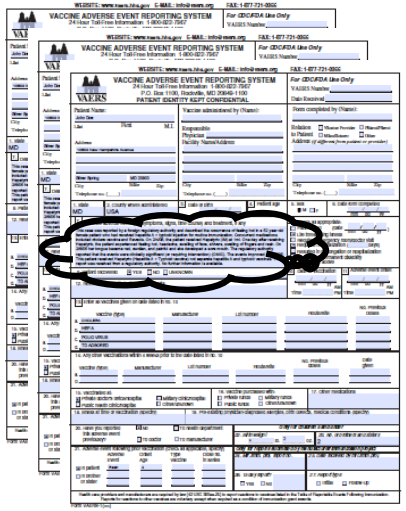
Bilateral weakness in lower extremities

Bilateral areflexia in lower extremities

CSF Mononuclear WBC count 3 cells/mm3; Protein 50 mg/dL

Electrodiagnostics Abnormal, consistent with polyneuropathy

Sounds like Guillain-Barré Syndrome



Medical Records

Data abstraction

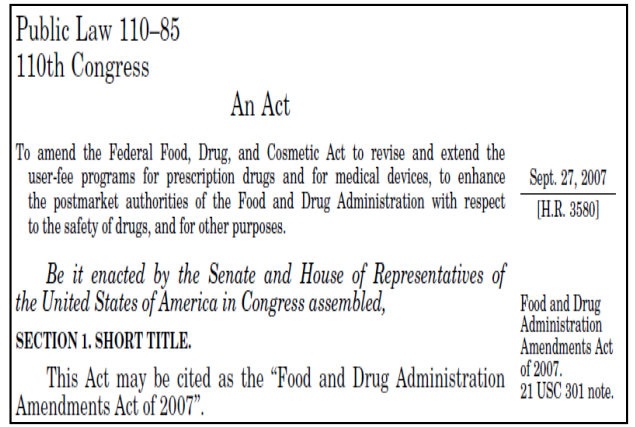
Study database

Case Definition

Expert Case Review

2007 FDA Amendments Act (FDAAA)

- Post Marketing Requirements
- Safety Labeling Changes
- Risk Evaluation and Mitigation Strategies (REMS)
- Required Safety Reviews ("915" and "921")
- **Active post-market Risk Identification and Analysis system**
  - **FDA Sentinel Initiative**



Public Law 110–85
110th Congress

An Act

To amend the Federal Food, Drug, and Cosmetic Act to revise and extend the user-fee programs for prescription drugs and for medical devices, to enhance the postmarket authorities of the Food and Drug Administration with respect to the safety of drugs, and for other purposes.

Sept. 27, 2007
[H.R. 3580]

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE.

This Act may be cited as the "Food and Drug Administration Amendments Act of 2007".

Food and Drug Administration Amendments Act of 2007.
21 USC 301 note.

Active Risk Identification and Analysis (ARIA) System

- Mandated creation in Section 905 of FDAAA 2007
- Linked to PMR in Section 901(3)(D)(i):
  - "The Secretary may not require the responsible person to conduct a study under this paragraph, unless the Secretary makes a determination that the reports under subsection (k)(1) and the **active postmarket risk identification and analysis system** as available under subsection (k)(3) will not be **sufficient** to meet the purposes set forth in subparagraph (B)."

Defining ARIA

- ARIA uses a subset of Sentinel System's full capabilities to fulfill the FDAAA mandate to conduct active safety surveillance

* Pre-defined, parameterized, and re-usable to enable faster safety surveillance in Sentinel (in contrast to protocol based assessments with customized programming)

† Electronic claims data, without manual medical record review

What is Sufficiency?

- Adequate data
  - Drug/biologic of interest and comparator
  - Confounders and covariates
  - Health outcome of interest
- Appropriate methods
- To answer the question of interest
  - assess a known serious risk related to the use of the drug/biologic
  - assess signals of serious risk related to the use of the drug/biologic
  - identify an unexpected serious risk when available data indicate the potential for a serious risk
- To lead to a satisfactory level of precision

When are automated queries insufficient?

- 43 Drug-AE pairs sufficient^
- 51 Drug-AE pairs insufficient^
- Reasons for Insufficiency*
  - Study population = 24
  - Exposure = 17
  - Outcome = 38
  - Covariate = 10
  - Analytic tool = 12

^1/2016-2/2018 – first 2 years of ARIA

*Total = 101 (some drug-AE pairs have more than one reason for insufficiency) - *preliminary results*

How do we improve sufficiency?

- Start “simple”
  - Add data partners (e.g. Medicare and HCA)
  - Create linkages (e.g. National Death Index and mother-infant)
  - Build new tools (e.g. Treescan for signal detection, distributed regression)
  - Add data to the Common Data Model (CDM) (e.g. physician specialty)

How do we improve sufficiency?

- When is “simple” not enough?
  - Outcomes with human expert-constructed algorithms, using data in the Sentinel CDM, resulting in insufficient PPV

Acute pancreatitis
Implant related complications (2)
osteosarcoma
Suicidal ideation and behavior
Opportunistic infections
Outpatient neutropenia

Stillbirth
Fluoroquinolone-associated disability
Neonatal enteroviral sepsis
Nerve injury
Anaphylaxis and serious hypersensitivity reactions (3)

How do we improve sufficiency?

- When is “simple” not enough?
  - Data not easily available for addition to CDM (e.g. lifestyle covariates in clinical narratives)
  - Data available but hard to standardize (e.g. laboratory, radiology, pathology results)
    - non-randomly missing so also need novel statistical methods
  - Cancer staging, severity, history, and therapeutic regimen

How do we improve sufficiency?

- When is “simple” not enough?
  - Direct linkage between claims and EHRs represents a small fraction of all patients in the Sentinel System
  - Many medical records only available as paper or PDF
  - 18 data partners in the Sentinel System don’t do everything exactly the same way

Might a machine-readable health record help?

- HOI algorithm identification and development
  - Apply machine learning to classified records to identify new algorithms from data already in CDM
  - Extract free-text fields from the machine readable health record, combined with claims, to create better algorithms
- Support epidemiologic studies via faster chart validation of outcomes, when a particular set of charts is needed

EHR narratives vs Coded data

Journal of the American Medical Informatics Association Advance Access published February 5, 2016

Extracting information from the text of electronic medical records to improve case detection: a systematic review

RECEIVED 12 May 2015
REVISED 13 October 2015
ACCEPTED 26 October 2015

Elizabeth Ford,[1] John A Carroll,[2] Helen E Smith,[1] Donia Scott,[2] and Jackie A Cassell[1]

ABSTRACT

Background Electronic medical records (EMRs) are revolutionizing health-related research. One key issue for study quality is the accurate identification of patients with the condition of interest. Information in EMRs can be entered as structured codes or unstructured free text. The majority of research studies have used only coded parts of EMRs for case-detection, which may bias findings, miss cases, and reduce study quality. This review examines whether incorporating information from text into case-detection algorithms can improve research quality.
Methods A systematic search returned 9659 papers, 67 of which reported on the extraction of information from free text of EMRs with the stated purpose of detecting cases of a named clinical condition. Methods for extracting information from text and the technical accuracy of case-detection algorithms were reviewed.
Results Studies mainly used US hospital-based EMRs, and extracted information from text for 41 conditions using keyword searches, rule-based algorithms, and machine learning methods. There was no clear difference in case-detection algorithm accuracy between rule-based and machine learning methods of extraction. Inclusion of information from text resulted in a significant improvement in algorithm sensitivity and area under the receiver operating characteristic in comparison to codes alone (median sensitivity 78% (codes + text) vs 62% (codes), $P = .03$; median area under the receiver operating characteristic 95% (codes + text) vs 88% (codes), $P = .025$).
Conclusions Text in EMRs is accessible, especially with open source information extraction algorithms, and significantly improves case detection when combined with codes. More harmonization of reporting within EMR studies is needed, particularly standardized reporting of algorithm accuracy metrics like positive predictive value (precision) and sensitivity (recall).

Keywords: electronic health records, review, text mining, data quality, case detection

INTRODUCTION

Information recorded in electronic medical records (EMRs), clinical reports, and summaries has the possibility of revolutionizing health-related research. EMR data can be used for disease registries, epidemiological studies, drug safety surveillance, clinical trials, and healthcare audits.

Information recording in EMRs

In most EMRs there is the possibility for the clinician both to code their findings in a structured format and also to enter information in narrative free text. There are various nomenclatures for structuring or coding information; the most widely used are International Classification of Diseases version 10,[1] Systematized Nomenclature of Medicine – Clinical Terms,[2] and the International Classification of Primary Care.[3] Within multi-modal EMRs there are also laboratory, pathology, and radiology reports, admission and discharge summaries, and chief complaints fields, which are in unstructured or semi-structured text. The balance of recording by the clinician, between codes and narrative text, is likely to vary by institution, EMR system, department, disease type, and component of the record.

Why do EMRs contain free text instead of being completely structured?

Clinicians experience a tension between choosing to code information and expressing it in text.[4] Among the main motivators for clinicians to code rather than use text is the increased ease of search, access, and retrieval.[5,6] A coded record allows the clinician to readily demonstrate that appropriate care has been provided, accurate diagnoses are made, and targets met.[7] This is especially important for billing after episodes of care, or for incentive based systems such as the National Health Service (NHS) Quality and Outcomes Framework in UK primary care.[8]

Coded data can be analyzed and summarized easily and on a large scale, whereas free text cannot. In contrast to structured data, narrative text is highly variable,[9] but is more engaging, captures the patient's narrative, can be told from different perspectives, and allows expression of feelings.[10] It is a better reminder for the clinician of the human encounter.[7]

Additionally, clinicians have given a number of reasons why they find coding onerous; the choices available in coded data may be too limiting, and may not allow for the expression of nuances.[11] The process of finding and entering codes on the computer represents an additional cognitive load,[5] and may take longer than summarizing the consultation in text.[6] Free text may be chosen when no code precisely describes clinical findings, or when there is a need to give supporting evidence for a diagnosis or suspicion.[12] Clinicians use free text as a pragmatic solution to recording vague diagnoses or strange collections of symptoms, when diagnoses need qualification, and for psychosocial problems.[7] Text can summarize processes of deduction, and modal language can be used to convey a range of possible outcomes.

Correspondence to Elizabeth Ford, Division of Primary Care and Public Health, Brighton and Sussex Medical School, Mayfield House, Village Way, Falmer, Brighton, BN1 9QH, UK; e.m.ford@bsms.ac.uk; Tel: (+44) 01273 641974. For numbered affiliations see end of article.

REVIEWS

Key Points

"**Text in EMRs** is accessible, especially with open source information extraction algorithms, and **significantly improves case detection when combined with codes**. More harmonization of reporting within EMR studies is needed, particularly standardized reporting of algorithm accuracy metrics like positive predictive value (precision) and sensitivity (recall)."

Authors also noted small sample that directly compared codes to narratives and **variability in performance.**

Ford E et al. J Am Med Inform Assoc 23 (5), 1007-1015. 2016.

Health Outcome of Interest: Anaphylaxis
Pilot project using OCR and NLP

- In the Sentinel System, most medical records only available as paper or PDF
- The human expert-constructed algorithm for anaphylaxis case identification has an “insufficient” PPV when using data in CDM
- Optical Character Recognition (OCR) of paper charts plus application of previously developed Natural Language Processing (NLP) and rule- and similarity-based algorithms for anaphylaxis case classification

Health Outcome of Interest: Anaphylaxis Mini-Sentinel claims-based algorithm

PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2013; 22: 1205–1213
Published online 5 September 2013 in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/pds.3505

ORIGINAL REPORT

Validation of anaphylaxis in the Food and Drug Administration's Mini-Sentinel

Kathleen E. Walsh[1]*, Sarah L. Cutrona[1,2], Sarah Foy[1], Meghan A. Baker[3,4], Susan Forrow[4], Azadeh Shoaibi[5], Pamala A. Pawloski[6], Michelle Conroy[7], Andrew M. Fine[8], Lise E. Nigrovic[8], Nandini Selvam[9], Mano S. Selvan[10], William O. Cooper[11] and Susan Andrade[1]

[1]Meyers Primary Care Institute, Worcester, MA, USA
[2]University of Massachusetts Medical School, Worcester, MA, USA
[3]Brigham and Women's Hospital, Boston, MA, USA
[4]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA
[5]Office of Medical Policy, CDER, FDA, Silver Spring, MD, USA
[6]HealthPartners Institute for Education and Research, Bloomington, MN, USA
[7]Massachusetts General Hospital, Boston, MA, USA
[8]Emergency Medicine, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA
[9]Government & Academic Research, HealthCore, Inc., Alexandria, VA, USA
[10]Comprehensive Health Insights, Humana Inc., Louisville, KY, USA
[11]Department of Pediatrics, Vanderbilt University, Nashville, TN, USA

ABSTRACT

Purpose We aim to develop and validate the positive predictive value (PPV) of an algorithm to identify anaphylaxis using health plan administrative and claims data. Previously published PPVs for anaphylaxis using International Classification of Diseases, ninth revision, Clinical Modification (ICD-9-CM) codes range from 52% to 57%.
Methods We conducted a retrospective study using administrative and claims data from eight health plans. Using diagnosis and procedure codes, we developed an algorithm to identify potential cases of anaphylaxis from the Mini-Sentinel Distributed Database between January 2009 and December 2010. A random sample of medical charts ($n = 150$) was identified for chart abstraction. Two physician adjudicators reviewed each potential case. Using physician adjudicator judgments on whether the case met diagnostic criteria for anaphylaxis, we calculated a PPV for the algorithm.
Results Of the 122 patients for whom complete charts were received, 77 were judged by physician adjudicators to have anaphylaxis. The PPV for the algorithm was 63.1% (95%CI: 53.9–71.7%), using the clinical criteria by Sampson as the gold standard. The PPV was highest for inpatient encounters with ICD-9-CM codes of 995.0 or 999.4. By combining only the top performing ICD-9-CM codes, we identified an algorithm with a PPV of 75.0%, but only 66% of cases of anaphylaxis were identified using this modified algorithm.
Conclusions The PPV for the ICD-9-CM-based algorithm for anaphylaxis was slightly higher than PPV estimates reported in prior studies, but remained low. We were able to identify an algorithm that optimized the PPV but demonstrated lower sensitivity for anaphylactic events.

KEY WORDS—anaphylaxis; serious allergic reaction; validation; administrative data; Food and Drug Administration; Mini-Sentinel; pharmacoepidemiology

Received 2 May 2013; Revised 18 July 2013; Accepted 26 July 2013

INTRODUCTION

In May 2008, the US Food and Drug Administration (FDA) launched the Sentinel Initiative, a long-term program designed to create a national electronic monitoring system for medical product safety.[1,2] The Mini-Sentinel pilot is a collaborative effort between the FDA and a consortium of institutions to develop the scientific operations needed for the eventual Sentinel System. An essential component of any active safety surveillance is the accurate and timely identification of adverse health outcomes. A key adverse health-related outcome of interest to the FDA and others is anaphylaxis.

*Correspondence to: K. E. Walsh, Meyers Primary Care Institute, Worcester, MA, USA. Email: kmackwalsh@yahoo.com

Key Points

- The authors developed and validated an algorithm using **administrative and claims data to identify cases of anaphylaxis.**
- The PPV for the overall algorithm was 63.1% (95% CI: 53.9-71.7%). **While this PPV improves on previous publications, it remains low.**
- The authors were able to identify an algorithm that optimized the PPV but demonstrated lower sensitivity for anaphylactic events.

Walsh KE et al. Validation of anaphylaxis in the Food and Drug Administration's Mini-Sentinel. Pharmacoepidemiology and drug safety 2013; 22: 1205–1213

Health Outcome of Interest: Anaphylaxis
VAERS NLP, Rule- and Similarity-based Classification

Research and applications

Vaccine adverse event text mining system for extracting features from vaccine safety reports

Taxiarchis Botsis,[1,2] Thomas Buttolph,[1] Michael D Nguyen,[1] Scott Winiecki,[1] Emily Jane Woo,[1] Robert Ball[1]

► Additional materials are published online only. To view these files please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2012-000881).

[1]Center for Biologics Evaluation and Research (CBER), Food and Drug Administration (FDA), Rockville, Maryland, USA
[2]Department of Computer Science, University of Tromsø, Tromsø, Norway

Correspondence to
Dr Taxiarchis Botsis, Office of Biostatistics and Epidemiology, CBER, FDA, Woodmont Office Complex 1, Room 306N, 1401 Rockville Pike, Rockville, MD 20852, USA;
taxiarchis.botsis@fda.hhs.gov

Received 3 February 2012
Accepted 28 July 2012

ABSTRACT

Objective To develop and evaluate a text mining system for extracting key clinical features from vaccine adverse event reporting system (VAERS) narratives to aid in the automated review of adverse event reports.

Design Based upon clinical significance to VAERS reviewing physicians, we defined the primary (diagnosis and cause of death) and secondary features (eg, symptoms) for extraction. We built a novel vaccine adverse event text mining (VaeTM) system based on a semantic text mining strategy. The performance of VaeTM was evaluated using a total of 300 VAERS reports in three sequential evaluations of 100 reports each. Moreover, we evaluated the VaeTM contribution to case classification; an information retrieval-based approach was used for the identification of anaphylaxis cases in a set of reports and was compared with two other methods: a dedicated text classifier and an online tool.

Measurements The performance metrics of VaeTM were text mining metrics: recall, precision and F-measure. We also conducted a qualitative difference analysis and calculated sensitivity and specificity for classification of anaphylaxis cases based on the above three approaches.

Results VaeTM performed best in extracting diagnosis, second level diagnosis, drug, vaccine, and lot number features (lenient F-measure in the third evaluation: 0.897, 0.817, 0.858, 0.874, and 0.914, respectively). In terms of case classification, high sensitivity was achieved (83.1%); this was equal and better compared to the text classifier (83.1%) and the online tool (40.7%), respectively.

Conclusion Our VaeTM implementation of a semantic text mining strategy shows promise in providing accurate and efficient extraction of key features from VAERS narratives.

INTRODUCTION

Spontaneous reporting systems (SRS), such as the vaccine adverse event reporting system (VAERS) play an important role in providing early evidence of new, serious, and unexpected adverse events after the use of medical products. In SRS, safety signals are typically identified using both qualitative and quantitative methods requiring time-intensive manual report review by medical experts. The standard approach to organizing the clinical data is to derive a description of the key features, including the diagnosis, time to onset, and alternative explanations that can be summarized across multiple cases for a 'case series' evaluation. Further evaluation of the summary data seeks unusual patterns among the key features or might include analysis of disproportionate reporting of a diagnosis after some medical products compared with others.[1] During the period 2006–11, the average number of reports per year more than doubled from the previous 5 years to 32 200; this trend of increasing numbers of reports to VAERS makes the development of automated tools essential.

To increase the efficiency of manual report review effectively, any automated tool must reliably: (1) recognize and differentiate between merely a reported symptom and an actual diagnosis in the narrative; (2) extract the key features that help determine whether a real association exists between a medical exposure and a reported outcome (eg, timing of both the exposure and outcome, alternative explanations for the event such as past medical history and co-administered medications; (3) reduce the amount of text required to be read and interpreted; and (4) tag and organize the key medical concepts after extraction from the raw report data.

A variety of methods has been employed to process medical text, extract facts, and/or recognize a specific range of adverse events in patient records, but each of these methods has certain limitations.[2–4] For example, machine learning techniques offer a promising solution but require large pre-annotated corpora for training, consuming considerable human resources.[5 6] Similarly, other approaches based on the construction of controlled dictionaries are considered to be laborious, demanding, and costly because they must be informed by specialist knowledge.[7] A number of text mining systems perform a part-of-speech-based tagging and shallow parsing that are followed by the named entity recognition, such as the Genia,[8] cTakes[6] and MedTAS/P systems.[9] MedLEE grammar combines semantic and syntactic co-occurrence patterns.[10] These systems do not have the capability to extract the key features required for safety surveillance using the case series framework without modification. While modification of one of these systems might be possible, we considered it would be simpler to develop a self-contained system. In this paper, we describe the development and evaluation of a text mining system specifically designed for VAERS that combines semantic tagging with rule-based techniques, to identify key clinical features and facilitate adverse event review.

Background

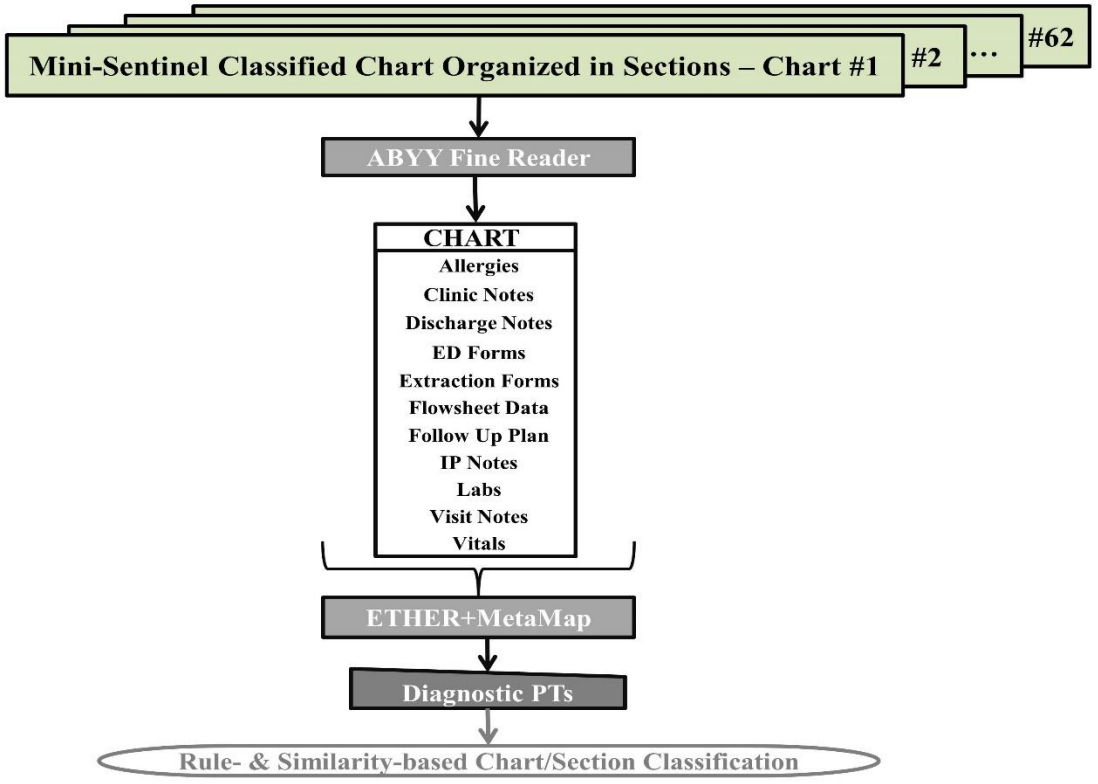
VAERS collects reports of adverse events following immunization (AEFI) with any US-licensed

Key Points

- The authors developed an algorithm to extract key features from narratives of Vaccine Adverse Event Report System (VAERS) reports using natural language processing.
- The authors used those features to classify reports of possible anaphylaxis after vaccination based on the Brighton Collaboration definition using both a rule-based and similarity-based classifier.

Botsis T, et al. Vaccine Adverse Event Text Mining (VaeTM) system for extracting features from vaccine safety reports. *J Am Med Inform Assoc* 19:1011-1018, 2012.

Health Outcome of Interest: Anaphylaxis Application of VAERS algorithm to MS charts



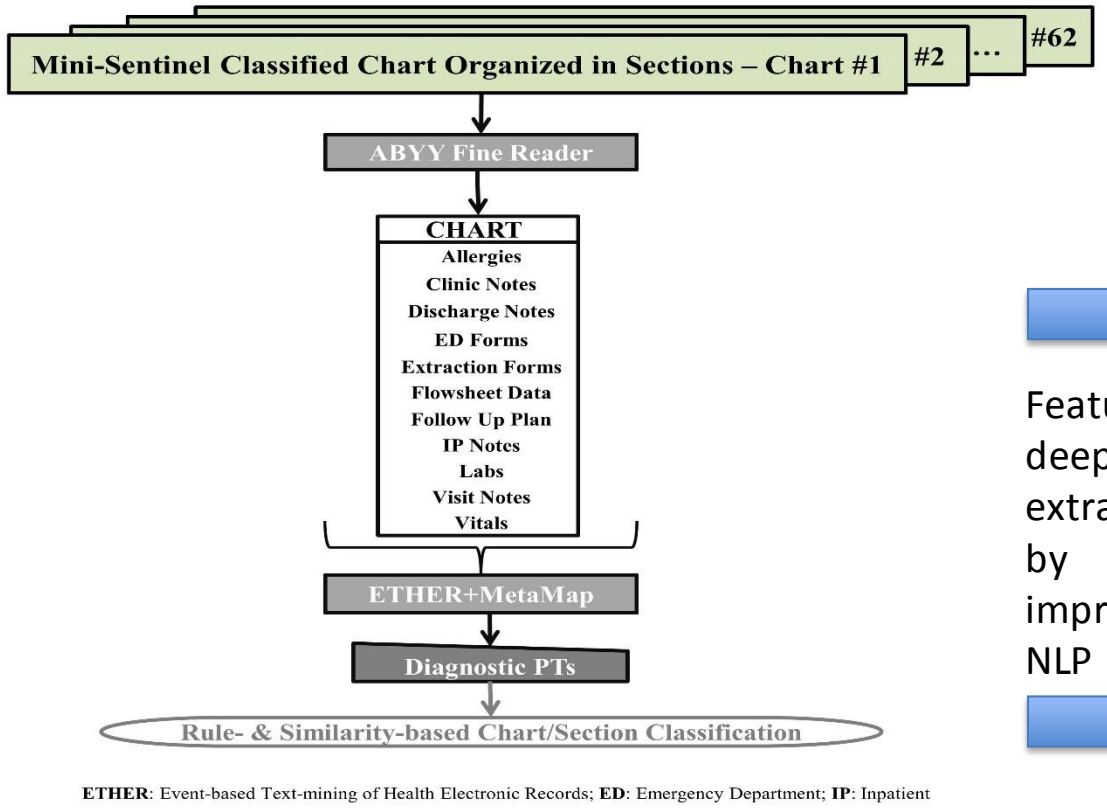
ETHER: Event-based Text-mining of Health Electronic Records; **ED**: Emergency Department; **IP**: Inpatient

Key Points

- The previously developed natural language processing, rule- and similarity-based classification approaches demonstrated almost equal performance (F-measure: 0.753 vs. 0.729, recall 100% vs 100%, precision 60.3% vs 57.4%).
- These algorithms might improve recall but had **similar precision (PPV) to claims only algorithms from MS**.

Ball et al, Evaluating automated approaches to anaphylaxis case classification using unstructured data from the FDA Sentinel System, under review

Health Outcome of Interest: Anaphylaxis Application of VAERS algorithm to MS charts



ETHER: Event-based Text-mining of Health Electronic Records; **ED**: Emergency Department; **IP**: Inpatient

Features for deep extraction by improved NLP

Key Points

- Reasons for misclassification included: the **inability** of the algorithms **to make the same clinical judgments as human experts** about the timing, severity, or presence of alternative explanations; the identification of terms consistent with anaphylaxis but present in conditions other than anaphylaxis.

Ball et al, Evaluating automated approaches to anaphylaxis case classification using unstructured data from the FDA Sentinel System, under review

Additional Challenges

- Solutions need to be implementable in a distributed data network
  - adaptable to run on native databases with very different formats
  - account for likely performance differences in different settings
- Potentially has implications for data governance and privacy preservation

Summary

- Many efforts to improve ARIA sufficiency underway
- “Non-simple” problems related to outcome validation might benefit from new technologies, such as NLP and machine learning
- Goal for workshop is to brainstorm ideas for 1, 3, 5 year projects and to know what to put into the 10 year bucket
- Solutions for improving ARIA sufficiency will likely also contribute to building Sentinel as a national resource for the learning healthcare system

Acknowledgements

- Michael Nguyen, Steve Anderson, Gerald Dal Pan, and Sentinel Team
- Jeff Brown, Judy Maro, Rich Platt, Sentinel Operations Center staff, and Sentinel Partners
- Adam Aten, Greg Daniel, Mark McClellan, and Duke Margolis staff

Thank you

Summary of Ongoing Projects and New Directions

Duke-Robert J. Margolis, MD, Center for Health Policy: Next Steps to Advance the Sentinel System

July 26, 2018

Jeffrey Brown, PhD

DEPARTMENT OF POPULATION MEDICINE

HARVARD MEDICAL SCHOOL

Harvard Pilgrim Health Care Institute

Projects in Progress

Project
Data Partner Data Assets and Expertise Survey
Data Partner Technical Assessment: Discovery and Planning
Data Sharing Guidance for Limited Datasets, patient profiles, and chart re-use
HOI 1.0 Validation (Serious Infections)
HOI 1.0 Validation (Lymphoma)
HOI 1.0 Validation (Stillbirth)

Projects Slated to Start / Planned

Project
HOI 2.0 Validation (Anaphylaxis)
HOI 2.0 Validation (Acute Pancreatitis)
MITRE CASAE engagement to assess new technologies for distributed networking
Chart Review Re-Engineering -Development of Chart Review Resource Intensity Score -Standardized SOPs for Chart Review -Discovery Phase for facility and provider SCDM fields
Vertical Distribute Regression Demonstration with CMS and PCORI sites

Opportunities for Improving the Efficiency of Outcome Validation in the Sentinel System

Project Categories

1. Chart Review Improvement Activities
   - Includes laying the groundwork for later HOI 2.0 methods
2. Common Data Model Readiness for Expansion
3. Methods Activities
4. Sentinel Patient Identifier and Linkage Activities

Chart Review Improvement Opportunities

Data Sources	Proposal Description
Charts	Scan Charts: Develop process to routinely scan charts at scale using optical character recognition tools
CDM	Improve Case Classification: Using existing Common Data Model data to develop machine learning methods to improve case classification (requires validated cases for learning)
EHR, CDM	Use corpus of validated cases and machine learning to assess whether claims data alone, claims + structured EHR, claims + unstructured EHR best identify cases

Chart Review Improvement Opportunities: Issues to consider

- Identify production-level Optical Character Recognition software and assess implementation barriers
  - Data storage, privacy, access, costs
- Address legal and regulatory issues with re-use of existing charts for other public health activities
- Assess potential to amend the “Dear Healthcare Provider” letters to allow for multiple uses of charts and chart-derived data

Common Data Model Infrastructure Opportunities

Data Sources	Proposal Description
CDM, EHR	Assess governance barriers and feasibility of populating CDM with unstructured free text notes
CDM	Add Sentinel and non-Sentinel funded chart validation information (ie, case status) to Common Data Model
CDM	Assess barriers to using charts obtained for other reasons (e.g. audits) to populate the Common Data Model with chart-extracted information
EHR	Evaluate value of EHR-only datasets for claims-compatible algorithm development

Common Data Model Infrastructure Opportunities: Issues to consider

- Ongoing data assets and methods expertise survey will inform data discovery projects for enhancement of the Common Data Model
  - Some partners may have data that could be incorporated into the data model quickly
- Expect substantial regulatory and legal hurdles related to re-use of chart-derived data
- Use of standardized versus unstructured information for rapid querying
  - Use of unstructured data requires time to make usable
  - Issues with patient privacy with unstructured data

Methods and Other Opportunities

Proposal Description
Implement machine learning for causal inference (ie, substitute investigator-driven propensity score model with machine learning methods)
Develop methods to use Missing Not-at-Random (MNAR) data; example: laboratory data values
Adapt doubly robust causal inference methods to a distributed database
Develop a process for rapid late-binding QA (example: lab data)
Partner with Health Information Exchanges to allow for rapid, focused chart retrieval
Develop alternatives to SAS-based querying infrastructure

Methods and Other Opportunities: Issues to consider

- Methods projects (e.g., Missing Not-at-Random information, doubly robust causal inference) require workgroup creation and appropriate data
- Regulatory, legal, and technical issues with working with Health Information Exchanges
- Software and technical barriers for using alternative to SAS-based distributed querying
  - Positive experience with PCORnet can be leveraged

Sentinel Patient Identifier and Linkage Opportunities

Data Sources	Proposal Description
CDM	Develop Sentinel Patient ID to identify same person across sites; assess overlap and proportion with enrollment transitions between existing partners
CDM, EHR	Demonstrate vertical distributed regression between sites to supplement claims data
CDM, EHR	Create a pilot claims-EHR linkage between Sentinel and PCORnet Data Partners

Classes of Methods for Linkage

- Identifiable: Use direct identifiers (like health information exchanges) or clear text identifiers
  - Example: PCORnet ADAPTABLE Clinical Trial. Patients will be individually consented for their participation anyway.
- Anonymized or Privacy Preserving Record Linkage (PPRL): Use anonymous hash identifier with secure transmission of the random seed (i.e. salt)
  - Personally Identifiable Information (PII) is converted into “tokens” and recombined using hashes and encryption
  - Could use trusted third party or exchange hash tables
  - Example: PCORnet Antibiotics Observational Study

EXTENDING COMPARATIVE EFFECTIVENESS RESEARCH AND MEDICAL PRODUCT SAFETY SURVEILLANCE CAPABILITY THROUGH LINKAGE OF ADMINISTRATIVE CLAIMS DATA WITH ELECTRONIC HEALTH RECORDS: A SENTINEL-PCORnet COLLABORATION

Prepared by: Kevin Haynes, PharmD, MSCE,[1] Nancy D. Lin, ScD,[2] Paul Avillach, MD, PhD,[3,4] Thomas W. Carton, PhD, MS,[5] Jeffrey R Curtis, MD, MS, MPH,[6] Kevin Fahey, MA,[7] Crystal Garcia, MPH,[8] Thomas Harkins, MA, MPH,[9] Wenke Hwang, PhD,[10] Cheryl N. McMahill-Walraven, MSW, PhD,[11] David Meltzer, MD, PhD,[12] Eliel Oliveira, MBA, MS,[5] Pamala A. Pawloski, PharmD,[13] Micah Prochaska, MD,[12] Jon Puro, MPA:HA,[14] Nandini Selvam, PhD, MPH,[1] Richard Platt, MD, MSc[8]

https://www.sentinelinitiative.org/sites/default/files/data/ComplementaryData/Sentinel_Sentinel-PCORnet-White-Paper_0.pdf

White Paper: Major Linkage Options

- Study-specific linkage
- Many-to-many linked dataset of identifiers to understand overlap
- Creation of a general purpose, persistent analyzable linked dataset

Take-home: Resolving governance policies is more challenging than technical challenges.

Summary

- Sentinel working on expanding analytic and surveillance capabilities across a range of areas
  - Chart Review Improvement Activities
  - Common Data Model Readiness for Expansion
  - Methods Activities
  - Sentinel Patient Identifier and Linkage Activities
- Regulatory, legal, and technical barriers exist
- Actively seeking partnerships with technology experts, new data sources, and other to address capability gaps

Closing Remarks