Sentinel

DATA-DRIVEN APPROACHES TO IMPROVE PHENOTYPE SENSITIVITY USING EHR DATA

Authors

Joshua C. Smith, PhD[1]; Daniel Park[1]; Jill Whitaker[1]; Michael F. McLemore, RN[1]; Elizabeth E. Hanchrow RN,MSN[1]; Dax Westerman[1]; Joshua Osmanski[1]; Robert Winter[1]; Saranrat Wittayanukorn, PhD[2]; Danijela Stojanovic, PharmD, PhD[2]; Arvind Ramaprasan[3]; Ann Kelley, MHA[3]; Mary Shea, M.A[3]; David J. Cronkite, MS[3]; Yueqin Zhao, PhD[2]; Darren Toh, ScD[4]; Kevin B. Johnson, MD, MS[5]; David Aronoff. MD[6]; David S. Carrell, PhD[3]

1Vanderbilt University Medical Center, 1211 Medical Center Dr, Nashville, TN 37232
2Division of Epidemiology, Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD
3Kaiser Permanente Washington Research Institute, 1730 Minor Ave, Seattle, WA 98101
4Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA
5Perelman School of Medicine at the University of Pennsylvania, 3400 Civic Center Blvd, Philadelphia, PA 19104
6Indiana University School of Medicine, 340 W 10th St, Indianapolis, IN 46202

ABSTRACT

- Medical product safety studies traditionally use condition-specific diagnosis codes as filters to identify patients with health outcomes of interest, but such filters may lack sensitivity.
- We sought to identify surrogate features for these diagnosis codes using electronic health record data (*coded* procedures, labs, medications, problem lists, and diagnoses) and evaluated whether such surrogates improved sensitivity by identifying cases overlooked by a traditional filter.
- Using EHR data from Vanderbilt University Medical Center (VUMC) and Kaiser Permanente Washington (KPWA), we identified a cohort of potential COVID-19 cases using six COVID-19-specific diagnosis codes as a traditional filter.
- The addition of EHR features increased true case sensitivity at VUMC and KPWA when identifying patients with both moderate and mild COVID-19, respectively.

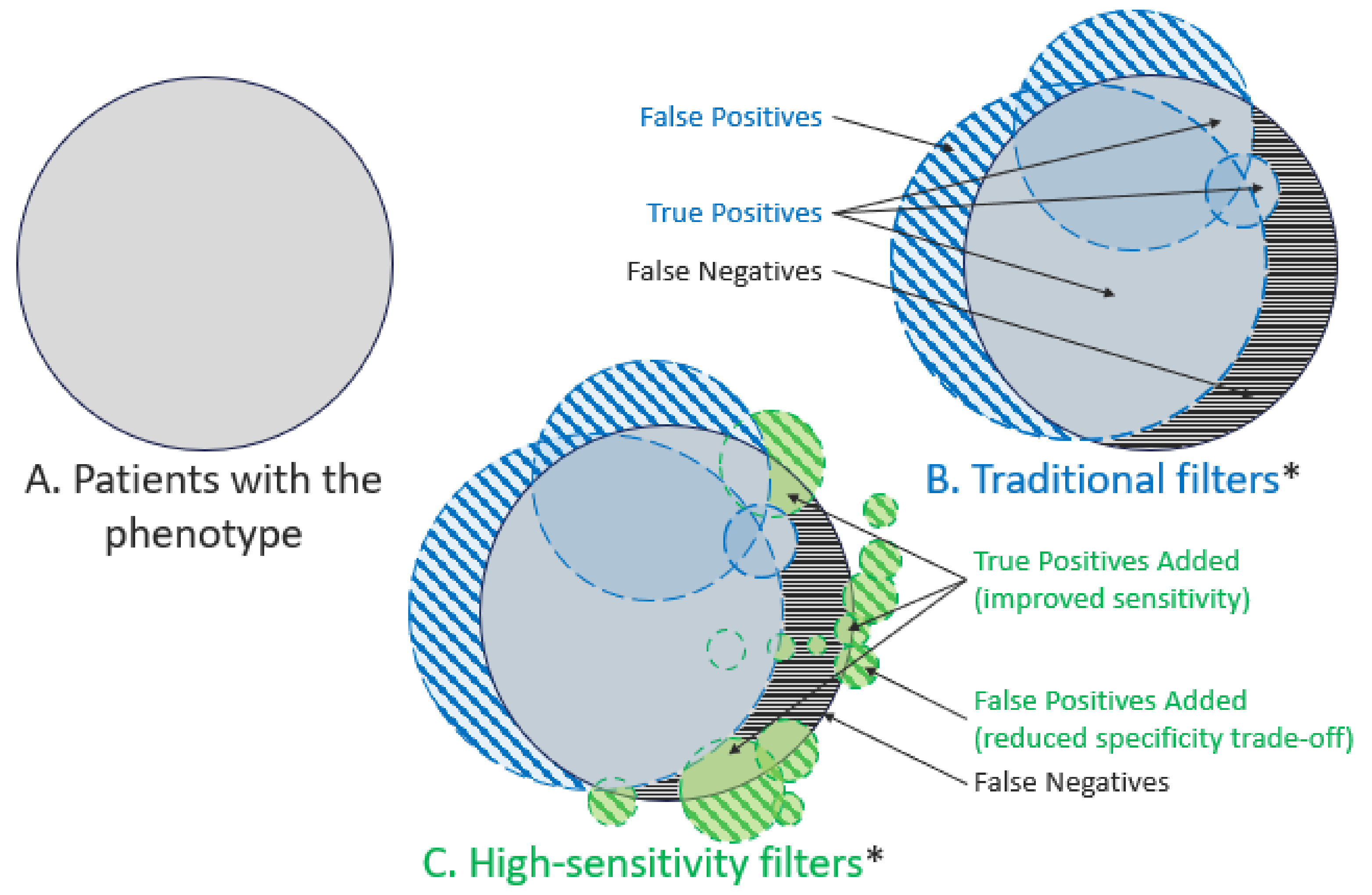
BACKGROUND

- Computable phenotype algorithms often use diagnostic codes as filters to identify presumptive cases (for which predictive models are developed to distinguish true cases from non-cases)
- Traditional filter examples
  - Anaphylaxis: ICD-10 codes T78.0*XA, T78.2XXA, T80.5*XA, T88.6XXA
  - Acute pancreatitis: ICD-9 code 577.0 and ICD-10 code K85
  - COVID-19: ICD-10 code U07.1
- But some true cases may not be coded with disease-specific diagnosis code, resulting in reduced sensitivity

OBJECTIVE

To improve identification of patients with a phenotype of interest (A), over and above traditional filtering approaches that identify presumptive cases using small sets of disease-specific diagnosis codes that are reasonably specific but overlook some true cases (B), via a data-driven approach for discovering other coded features that may serve as surrogates for traditional filters (C), thereby improving overall sensitivity at reasonable cost in specificity.

We illustrate application of the method by applying it to a COVID-19 phenotype.



False Positives
True Positives
False Negatives
A. Patients with the phenotype
B. Traditional filters*
True Positives Added (improved sensitivity)
False Positives Added (reduced specificity trade-off)
False Negatives
C. High-sensitivity filters*

* These figures provide hypothetical examples and are not representative of our COVID-19 phenotype.

METHODS

Rationale: Structured data features occurring much more commonly near a traditional filter (e.g., U07.1 “COVID-19 disease”) may serve as useful surrogates.

Process: Identifying coded healthcare data that may supplement traditional filters for identifying presumptive patients:

1. Identify a universe of patients among which you want to identify patients with the phenotype of interest (COVID-19: adults with ≥1 visit during 4/1/2020-3/31/2021)
2. Specify the traditional filter code (COVID-19: ICD code *U07.1 “COVID-19 disease”*)
3. Identify all candidate high-sensitivity filter codes. These include:
   - Any code appearing within +/- 3 days of a visit coded with a traditional filter code
   - Any code type (diagnosis, procedure, medication, lab, or problem list entry)
   - Note: Each code type may yield scores to hundreds of unique candidate codes
4. For each candidate code, compute a *relative ratio (RR)* indicating how much more commonly the candidate code appears near visits with a traditional filter code than visits without a traditional filter code:

$$\mathrm{RR} = \frac{\left(\frac{\textit{Number of patients with a candidate code} \leq 3 \textit{ days from a visit with the traditional code}}{\textit{Number of patients with the traditional code in any visit}}\right)}{\left(\frac{\textit{Number of patients with a candidate code} > 3 \textit{ days from a visit with the traditional code}}{\textit{Number of patients without the traditional code in any visit}}\right)}$$

5. Calculate *number of new patients* each candidate code would add over and above those identified by the traditional filter
6. Exclude codes with *RR<10 at either study site*
7. Manually review remaining codes and retain those that:
   a) Have clinical face validity and
   b) Do not add very large numbers of new patients (≤ moderate increase in sample)

RESULTS

A total of 749,353 VUMC and 717,379 KPWA adult patients has ≥1 visit during the study period, of which 23,388 VUMC and 17,398 KPWA patients had at least one encounter with a traditional COVID-19 disease-specific diagnosis code. We used encounters within +/-3 days of these COVID-19-coded encounters to identify candidate features to serve as surrogate/high-sensitivity filters.

Table 1 summarizes candidate diagnoses, procedures, medications, labs, and EHR problem list entries identified and analyzed as candidate high-sensitivity filters. Based on clinician review of candidate features we selected a total of 43 features for use as high-sensitivity filters and used this set at both study sites. No lab codes had sufficient clinical face validity to serve as surrogates for the traditional COVID-19 filter.

Table 1. Summary of structured data features considered as surrogates for identifying patients with symptomatic COVID-19 disease by discovery stage (rows), code type (columns), and study site (VU=VUMC, KP=KPWA, columns).

Discovery stage	Diagnoses		Procedures		Medications[1]		Labs		Prob. List	
	VU	KP	VU	KP	VU	KP	VU	KP	VU	KP
Candidate features (within +/- 3 days of a visit with U07.1)	33,187	30,803	14,683	15,824	1,913	17,190	1,922	132	NA[2]	11,535
Features manually reviewed (RR ≥10)	70	163	37	209	24	252	3	0	NA[2]	28
Features selected for use in both sites	24		10		4		0		5	
Total features	43									

[1] VUMC medications included *inpatient* and *outpatient* data by *ingredient*; KPWA data were *outpatient* by National Drug Code (NDC).
[2] We did not use VUMC EHR Problem list data

Illustrative high-sensitivity filter codes with *RR≥10* at *either study site*

- Diagnosis code J12.89 “Other viral pneumonia”
  VUMC RR=877, added potential pts=37; KPWA RR=821, added potential pts=38
- Procedure code XW033E5 “Remdesivir Anti-infective into Central Vein”
  VUMC RR=328, added potential pts =5; KPWA RR=25,584, added potential pts=1
- Medication name “Baricitinib”
  VUMC RR=55, added potential pts =10; KPWA RR=0, added potential pts=5
- Problem list entry “Acute respiratory distress syndrome”
  VUMC RR=8, added potential pts =128; KPWA RR=37, added potential pts=60

Added potential COVID-19 patients and estimated added actual cases identified by all 43 high-sensitivity filter codes are summarized in Table 2.

- Sensitivity increased ~12% (with an increase in potential patients of ~22%)

Table 2. Potential COVID-19 patients identified by a traditional filter (ICD-10 code U07.1 “COVID-19 disease”) and added potential patients identified by high-sensitivity filters (any of 43 in Table 1) with estimated actual cases added, by site.

Study site	Traditional filter		High-sensitivity filters	
	Potential patients	Estimated actual cases*	Added potential patients (increase)	Estimated added actual cases* (increase)
VUMC	20,951	18,856	4,566 (+22%)	2,511 (+13%)
KPWA	6,847	4,861	1,482 (+22%)	563 (+12%)

* We estimated the number of actual cases based on manual chart reviews of random samples of potential patients.

CONCLUSION

Identification of structured data surrogates for traditional, disease-specific diagnosis codes to improve sensitivity of identifying patients who may have a phenotype of interest:

- Can be done by a phenotype-independent, data-driven, semi-automated approach
- Improved sensitivity for patients with COVID-19 disease by 12%-13% with an acceptable increase in overall sample size
- May be implemented via publicly-available parameterized SAS® programs available at https://github.com/kpwhri/Sentinel-Scalable-NLP

LIMITATIONS

Limitations of this work include:

- To date it has only been applied to one phenotype
- Improvements in sensitivity may vary by phenotype

ACKNOWLEDGEMENTS/DISCLOSURES

- This project was supported by Task Order 75F40119F19002 under Master Agreement 75F40119D10037 from the U.S. Food and Drug Administration (FDA).
- Many thanks are due to members of the Sentinel Innovation Center Workgroup that provided critical feedback during development of this work
- KBJ discloses affiliations with the Robert Wood Johnson Foundation and the University of Pennsylvania
- The views expressed in this presentation represent those of the presenters and do not necessarily represent the official views of the U.S. FDA.