## Research and Applications

# Adverse drug event rates in pediatric pulmonary hypertension: a comparison of real-world data sources

**Alon Geva,[1,2,3] Steven H. Abman,[4,5] Shannon F. Manzi,[1,6,7] Dunbar D. Ivy,[5,8] Mary P. Mullen,[7,9] John Griffin,[2] Chen Lin,[1] Guergana K. Savova,[1,7] and Kenneth D. Mandl[1,7,10]**

[1]Computational Health Informatics Program, Boston Children's Hospital, Boston, Massachusetts, USA, [2]Division of Critical Care Medicine, Department of Anesthesiology, Critical Care, and Pain Medicine, Boston Children's Hospital, Boston, Massachusetts, USA, [3]Department of Anaesthesia, Harvard Medical School, Boston, Massachusetts, USA, [4]Division of Pediatric Pulmonary Medicine, Children's Hospital Colorado, Aurora, Colorado, USA, [5]Department of Pediatrics, University of Colorado School of Medicine, Aurora, Colorado, USA, [6]Division of Genetics & Genomics, Clinical Pharmacogenomics Service, Department of Pharmacy, Boston Children's Hospital, Boston, Massachusetts, USA, [7]Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, USA, [8]Division of Cardiology, Heart Institute, Children's Hospital Colorado, Aurora, Colorado, USA, [9]Department of Cardiology, Boston Children's Hospital, Boston, Massachusetts, USA, and [10]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA

Corresponding Author: Alon Geva, MD, MPH, Division of Critical Care Medicine, Department of Anesthesiology, Critical Care, and Pain Medicine, Bader 634, Boston Children's Hospital, 300 Longwood Avenue, Boston, MA 02115, USA (alon.geva@childrens.harvard.edu)

### ABSTRACT

**Objective:** Real-world data (RWD) are increasingly used for pharmacoepidemiology and regulatory innovation. Our objective was to compare adverse drug event (ADE) rates determined from two RWD sources, electronic health records and administrative claims data, among children treated with drugs for pulmonary hypertension.

**Materials and Methods:** Textual mentions of medications and signs/symptoms that may represent ADEs were identified in clinical notes using natural language processing. Diagnostic codes for the same signs/symptoms were identified in our electronic data warehouse for the patients with textual evidence of taking pulmonary hypertension-targeted drugs. We compared rates of ADEs identified in clinical notes to those identified from diagnostic code data. In addition, we compared putative ADE rates from clinical notes to those from a healthcare claims dataset from a large, national insurer.

**Results:** Analysis of clinical notes identified up to 7-fold higher ADE rates than those ascertained from diagnostic codes. However, certain ADEs (eg, hearing loss) were more often identified in diagnostic code data. Similar results were found when ADE rates ascertained from clinical notes and national claims data were compared.

**Discussion:** While administrative claims and clinical notes are both increasingly used for RWD-based pharmacovigilance, ADE rates substantially differ depending on data source.

**Conclusion:** Pharmacovigilance based on RWD may lead to discrepant results depending on the data source analyzed. Further work is needed to confirm the validity of identified ADEs, to distinguish them from disease effects, and to understand tradeoffs in sensitivity and specificity between data sources.

**Key words**: adverse drug event, administrative claims, healthcare, natural language processing, hypertension, pulmonary

## INTRODUCTION

Measurement of medication-related adverse effects is a critical aspect of drug evaluation. Traditional pharmacovigilance relies on varied data sources, including randomized clinical trials (RCTs), observational studies, spontaneous reports such as those collected in the Food and Drug Administration's Adverse Event Reporting System (FAERS, commonly referred to as MedWatch), and manual chart review of data in electronic health records (EHRs).[1] Although drug manufacturers are required to submit postmarket adverse event reports to the Food and Drug Administration (FDA), this information is not uniformly available to clinicians.[2] The 21st Century Cures Act directs the FDA to use real-world data (RWD) in the drug approval process. Generally, RWD refers to any data generated outside of clinical trials but most often is used to designate data produced in the course of routine delivery of healthcare.[3] Use of RWD is particularly important for medications that are commonly used off-label, such as those targeted for treatment of pulmonary hypertension (PH) in children.[2] Insurance claims are at the core of systems for RWD-based pharmacovigilance, such as the FDA's Sentinel program, because insurance claims reliably provide longitudinal data about medication dispensing and medical findings.[4] However, such structured data for identifying diagnoses may lack sensitivity for detecting adverse drug events (ADEs), as not all signs and symptoms are recorded for billing purposes.[5]

Clinical notes in EHRs provide an alternative information source for ADE detection.[1] An increasing number of studies are examining the use of natural language processing (NLP) to extract ADE knowledge from the clinical narrative,[1,6] social media posts,[7] and reporting databases like MedWatch.[8] Investigators have examined the relative value of structured and unstructured EHR data for detecting disease.[9–12] We sought specifically to compare 2 common approaches to ADE ascertainment from RWD—1 using an NLP pipeline to identify potential ADEs in free text clinical notes, and another using diagnostic codes. We compare potential ADE rates ascertained from diagnostic codes both within our EHR, which includes all clinician-indicated diagnostic codes, and from claims to a national, private health plan. We apply these approaches to studying ADE rates for PH-targeted drugs in children. The focus on medications for pediatric PH is particularly informative, as these drugs are uniformly used off-label in children, and their safe and efficacious use is primarily supported by data from adult studies.[2]

## MATERIALS AND METHODS

### Data sources

The study populations were created from 2 retrospective data sources. The first, which we refer to as the EHR dataset, was created as part of a parent project studying pediatric pulmonary vascular disease. The pediatric PH cohort was identified from the Boston Children's Hospital data warehouse through a computable phenotype, which has an 85% positive predictive value (PPV) for identifying patients with pediatric PH.[13] The original patient cohort included adult patients with childhood-onset PH. For the current study, we excluded patients who were 20 years of age and older. We extracted plain-text admission, discharge, consultation, progress, emergency department, procedure, and clinic notes for these patients. We also obtained diagnostic codes from the local Informatics for Integrating Biology and the Bedside (i2b2) data warehouse,[14] which is a research "sidecar" to the EHR that captures, among other clinical data, clinician-entered coded diagnoses[15] for these same patients.

The second cohort, which we refer to as the claims dataset, was derived from a national, private health plan in the United States. The dataset included claims for approximately 70 million beneficiaries filed from January 2008 to February 2016. Available data included beneficiary demographics (sex, age), dates, and prescription details for dispensed medications, and dates and diagnostic codes associated with inpatient and outpatient hospital visits.

Relevant medication and ADE pairs of interest were identified *a priori* based on review of the literature[16–22] and complemented by input from members of the Pediatric Pulmonary Hypertension Network (PPHNet) and the National Heart, Lung, and Blood Institute Pediatric Pulmonary Vascular Disease Outcomes Bioinformatics Clinical Coordinating Center Investigators. Signs/symptoms representing potential ADEs were grouped based on similar pathophysiology and terminology (Table 1). All formulations of a particular drug were analyzed together. The following medications were used in cohort definitions to maximize sensitivity: ambrisentan, bosentan, epoprostenol, iloprost, macitentan, riociguat, treprostinil, sildenafil, or tadalafil. Given that treprostinil and epoprostenol infusions are often initiated inpatient, and thus may not be represented in pharmacy benefits claims data, they were not included in the ADE analysis. We also excluded other formulations and dose forms of these medications given their rare use in both datasets. Similarly, iloprost, macitentan, and riociguat are rarely used at our institution (and the latter 2 are rarely represented in the claims data) and, thus, were also excluded from the ADE analysis. The study was approved by the institutional review board at Boston Children's Hospital with exemption from review for analysis of claims data and waiver of informed consent for review of EHR data.

### Clinical notes data

Notes for patients in the EHR dataset were processed using the Apache clinical Text Analysis Knowledge Extraction System (cTAKES),[23,24] an open-source system for clinical NLP. We used cTAKES to identify concept unique identifiers for Unified Medical Language System (UMLS)-based terms for medications and signs/symptoms of interest[25,26] as well as their attributes for negation, conditional status, and temporality relative to the document creation time.[23] A pair of a relevant medication and sign/symptom had to be temporally consistent to be considered a potential ADE. For example, a medication described with temporality *after* ("will start

**Table 1.** Adverse drug events considered in the study

- Anemia
- Diarrhea
- Edema
- Headache
- Hearing loss
- Dizziness/hypotension
- Intracranial hemorrhage
- Priapism
- Rash/flushing
- Reflux
- Seizure
- Sinusitis
- Syncope/pre-syncope
- Thrombocytopenia/bleeding
- Transaminitis
- Visual changes (including ischemic optic neuropathy)

sildenafil") could not cause a rash described with temporality *before* ("had a rash last week"). Medications with temporality *after* the document creation time were excluded.

Through iterative algorithm refinement, we defined a context of 25 newline characters for the candidate pairs. We further optimized the detection of negation and conditional phrases by adding cue words. We excluded certain terms that generated false-positive associations by subsuming them into a more specific concept (eg, specifying that "heparin flush" was not an instance of "flushing" as a sign/symptom). Thus, we retained non-negated, nonconditional mentions of medications and signs/symptoms of interest. The $F_1$ score—the harmonic mean of precision (positive predictive value) and recall (sensitivity)—computed on held out data (38 notes for 12 patients) was 0.78, with a precision of 0.69 and recall of 0.90 (interannotator agreement of 0.88 using Cohen's κ). Patients were considered to have been exposed to a medication if the medication was mentioned in at least 2 notes. Similarly, to decrease spurious detection of ADEs, a potential ADE was only counted if it was mentioned in at least 2 notes.

### EHR diagnostic codes

For the 263 patients taking a PH-targeted medication as identified based on clinical notes, we extracted all diagnostic (International Classification of Diseases, 9th revision [ICD-9] or 10th revision [ICD-10]) codes from the hospital data warehouse. We created a dictionary of ICD-9 and ICD-10 codes for signs/symptoms that could represent potential ADEs (Supplementary Material Table S1) and used this dictionary to select potentially relevant diagnostic codes. We then excluded diagnostic codes that were entered prior to the first date of the first clinical note for a patient that mentioned the relevant PH-targeted medication. Finally, to maximize specificity, only diagnostic codes that appeared at least twice after the mention of a medication were considered a potential ADE. Algorithms using at least 2 care encounters related to a condition based on ICD codes have previously been validated for a range of diseases.[27,28]

### Payor claims data

Patients with pediatric PH were identified in the claims data as those under 20 years old having at least 2 claims for PH based on ICD-9 and/or ICD-10 codes.[29] We used the following ICD-9 codes to identify PH: 416.0, 416.8, 416.9, and 747.83. The ICD-10 codes used were I27.0, I27.2, I27.89, I27.81, I27.9, and P29.3. For completeness, we also included all patients under 20 years of age to whom a PH-targeted drug was dispensed, since these medications are used in children almost exclusively to treat PH. Of the 253 patients with claims for relevant medications 240 (95%) were also identified based on ICD-9 or ICD-10 codes; the remaining 13 patients all had a single diagnostic code for pulmonary hypertension.

Medications were identified in the claims data using National Drug Codes. All formulations of a medication were mapped to a single identifier based on the generic medication names associated with the National Drug Codes. We complemented the analytic approach above with a comparison to claims data, in part, because treatment duration could not be accurately identified in the EHR dataset. In the claims dataset, treatment episodes were defined based on the proportion of days covered (PDC), a method for measuring adherence using claims data.[30–32] In brief, for each patient–medication dyad, the first treatment episode was identified commencing with the date indicated by a claim for that medication being dispensed. Based on the days supplied indicated in the claim, the treatment episode was extended until a gap in medication availability was identified (Figure 1). In order to focus on the subset of the cohort with high PDC (ie, high medication adherence), we used the upper whisker of the Tukey boxplot[33] to define outlier treatment gaps. Using this criterion, treatment gaps without medication supply longer than 25 days represented separate treatment episodes. Stop dates for each treatment episode were considered the last date of medication availability in the treatment episode. PDC was calculated for the resulting treatment episodes; treatment episodes with PDC less than 0.8 were excluded from analysis. Thus, signs/symptoms corresponding to potential ADEs that were billed on dates within treatment episodes likely occurred while patients were exposed to the medications of interest. In sensitivity analyses, we examined the effect of decreasing the gap in medication supply used to define treatment episodes to the 75th percentile of medication supply gaps, or 4 days; varying the PDC cutoff for treatment periods from 0.7 to 0.9; and considering the treatment episode to extend up to 60 days after the last medication supply day.

We used the same dictionary of diagnostic codes for potential ADEs used for the EHR dataset. Only signs/symptoms that occurred during the time interval in which the patient was on a medication were considered. Similar to our approach for calculating ADE rates from clinical notes, only pairs of medications and signs/symptoms that occurred at least twice for a patient were considered potential ADEs. The number of patients at risk for the ADE was calculated as the number of patients with at least 2 claims for a medication.

### Analysis

Prevalence of a potential ADE was calculated as the number of patients identified as having experienced the relevant sign or symptom divided by the number of patients potentially at risk. Patients at risk of each ADE were defined as those who were exposed to a medication suspected of causing that ADE. Number of patients exposed to each drug was determined based on non-negated, nonconditional mentions of the medication in the clinical notes for the EHR dataset or claims for those medications in the claims dataset. The 95% confidence intervals were calculated using the binomial distribution. The relative rate of potential ADEs detected in each dataset was calculated as the ratio of the 2 prevalence estimates, with 95% confidence intervals calculated using the normal approximation. No relative rate or confidence interval was calculated when a potential ADE was identified in only 1 dataset. Demographic variables were compared using Student's *t* test for continuous variables or Pearson's chi-squared test for categorical variables. The nonparametric bootstrap procedure in the *MRCV* package[34] was used to compare the number of patients exposed to medications between datasets. A *P* value < .05 was considered statistically significant. All data preparation and analysis was performed using R version 3.5.0.[35]

## RESULTS

### Data sources

Of 982 patients seen at Boston Children's Hospital who were found to have at least transient PH,[13] 286 (29%) were found to have used PH-targeted medications based on analysis of their clinical notes. Twenty-three of these patients were only prescribed excluded PH-targeted medications and were not included in the final cohort.

The claims dataset included 6,233 beneficiaries under 20 years old with at least 2 diagnostic codes for PH or claims for PH-targeted medications, of whom 253 (4.1%) had claims for at least 1 PH-targeted drug. All patients with PH-targeted medication claims had
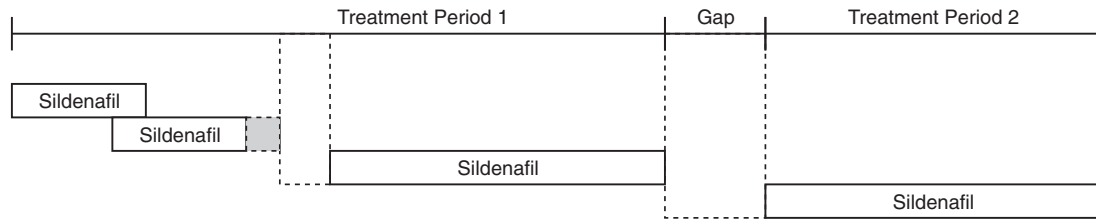
**Figure 1.** Schematic of how treatment periods were constructed for the claims dataset. Boxes represent claims for dispensed medications, such as sildenafil. The shaded box represents adjustment of the available medication days to account for the overlapping medication fill in a manner analogous to calculation of the proportion of days covered (see Methods for details). In this example, the shorter gap in medication availability during treatment period 1 falls below the threshold for determining start and end dates for medication exposure. Thus, there are 2 treatment periods, separated by a longer gap in medication availability, during which the patient would be considered exposed to sildenafil.

**Table 2.** Characteristics of patients from electronic health record (EHR) and claims datasets

|  | EHR (N = 263)[a] | Claims (N = 253)[a] |
|---|---|---|
| Age (years)[b] | 3.7 ± 5.8 | 5.9 ± 6.4 |
| Male sex[c] | 136 (52%) | 127 (50%) |
| Year medication mentioned/filled[d] | 2014 (2012–2015) | 2013 (2011–2013) |
| Medication prescribed[e] |  |  |
| Sildenafil | 252 (96%) | 217 (86%) |
| Tadalafil | 35 (13%) | 38 (15%) |
| Bosentan | 47 (18%) | 46 (18%) |
| Ambrisentan | 15 (5.7%) | 26 (10%) |

[a]Details are shown only for patients prescribed at least one PH-targeted medication of interest.

[b]Age at first mention or filling of PH-targeted medication (mean ± standard deviation).

[c]Frequency (percent).

[d]Median (interquartile range).

[e]Frequency (percent); sum is greater than 100% due to patients prescribed multiple medications.

claims for at least 1 of the medications included in the ADE analysis. Instances of PH-targeted drug exposure occurred during slightly later years in the EHR dataset as compared to the claims dataset ($P < .001$; Table 2). Patients in the claims dataset were older ($5.9 \pm 6.4$ y vs $3.6 \pm 5.7$ y, $P < .001$). In both datasets, sex was similar ($P = .41$), and sildenafil was the most commonly prescribed drug. More patients in the EHR dataset were prescribed sildenafil, whereas fewer were prescribed ambrisentan ($P = .003$).

## Comparison of potential ADE rates found in EHR clinical notes versus diagnostic codes

Rates of many potential ADEs differed between the EHR clinical notes and diagnostic codes (Figure 2). Of 40 potential ADEs examined, 6 (15%) were identified significantly more frequently in the EHR clinical notes. An additional 13 potential ADEs were identified only in clinical notes but not in diagnostic codes. Only 1 potential ADE—hearing loss associated with sildenafil—was identified significantly more frequently in the diagnostic codes. Potential ADE rates for each PH-targeted drug based on analysis of free-text clinical notes in the EHR dataset are shown in Figure 3A. Figure 3B shows potential ADE rates obtained from the diagnostic codes. Of note, rates for some potential ADEs were similar in both data sources. For instance, gastroesophageal reflux (GER) associated with sildenafil use was the most commonly identified potential ADE, present in
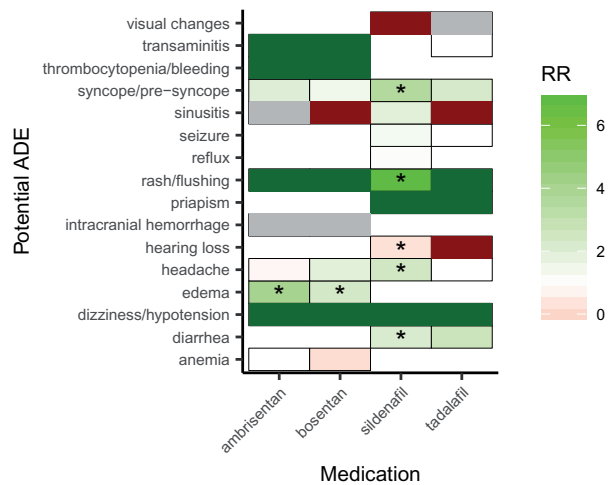


**Figure 2.** Relative rates of adverse drug events (ADEs) found in EHR clinical notes versus EHR diagnostic codes. Cells corresponding to medication-ADE pairs that were more frequent in the EHR clinical notes are shaded green, whereas those more frequent in the EHR diagnostic codes are shaded red. Darker colors indicate higher relative frequency, and solid colors indicate medication-ADE pairs found exclusively in one dataset. Gray cells indicate medication-ADE pairs that were found in neither dataset. Asterisks indicate medication-ADE pairs whose relative rate between the 2 datasets was significantly greater than or less than 1.

47% (95% confidence interval (CI), 41%–53%) of patients based on clinical notes and 42% (95% CI, 36%–49%) of patients based on diagnostic codes.

We also found differences in ADE rate ascertainment from clinical notes versus the national health plan claims dataset. Similar to our findings within the EHR dataset, analysis of clinical notes generally identified more potential ADEs than diagnostic codes. Fourteen (35%) of 40 ADEs were found significantly more frequently in the free-text clinical notes, whereas only 1 ADE—again, hearing loss associated with sildenafil—was found more frequently using diagnostic codes (Figure 4). Again, similarly to the analysis within the EHR dataset, sildenafil-associated GER was the most commonly identified potential ADE, present at a similar rate (38% [95% CI, 31% – 44%]) in patients in the claims dataset as in the EHR dataset.

In sensitivity analyses, results changed only minimally when parameters for calculating treatment periods were modified to increase or decrease the sensitivity of ADE detection. When the minimum PDC was increased to 0.9, only 179 patients qualified as being on sildenafil, and the relative rate of syncope/presyncope with bosentan in the EHR dataset was no longer significantly greater
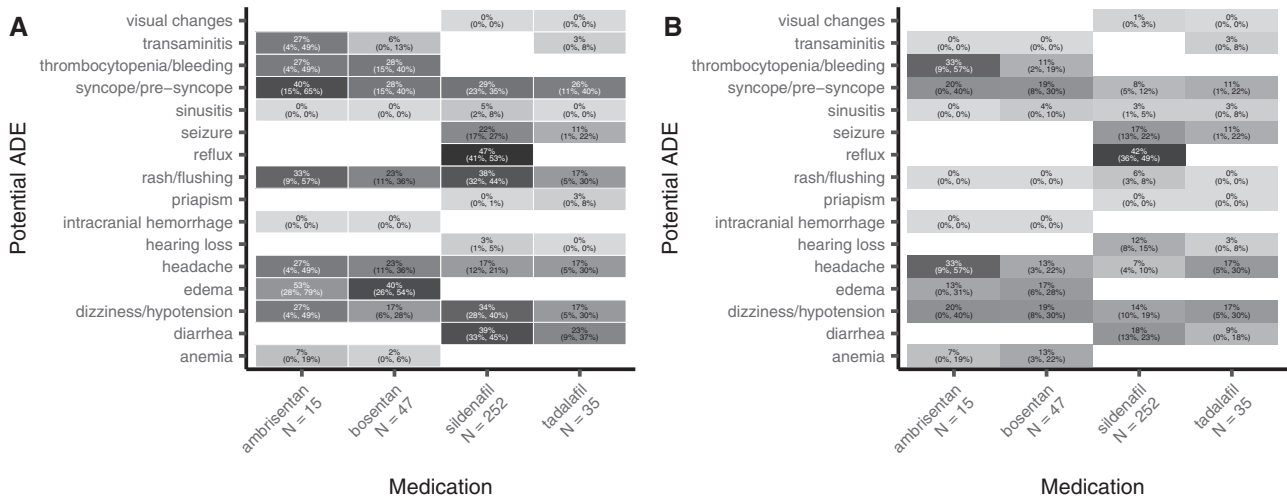
**Figure 3.** Rates of adverse drug events (ADEs) based on analysis of EHR clinical notes (A) and EHR diagnostic codes (B) data. Numbers within each cell indicate the percent (95% confidence interval) of patients on each medication experiencing a particular ADE. Darker cells indicate a higher frequency of the ADE.
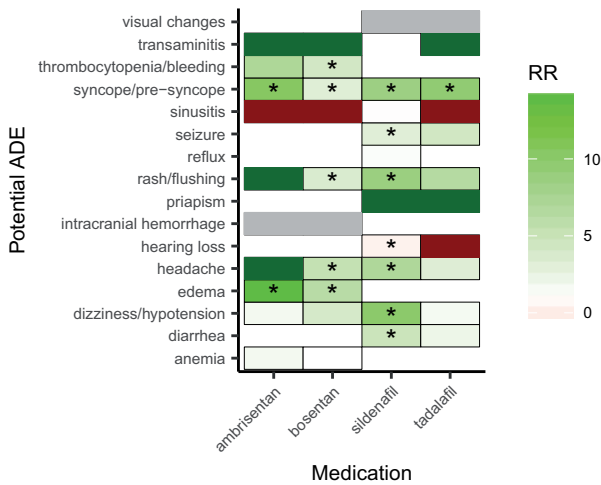


**Figure 4.** Relative rates of adverse drug events (ADEs) found in EHR clinical notes versus claims datasets. Cells corresponding to medication-ADE pairs that were more frequent in the EHR clinical notes are shaded green, whereas those more frequent in the claims are shaded red. Darker colors indicate higher relative frequency, and solid colors indicate medication-ADE pairs found exclusively in 1 dataset. Gray cells indicate medication-ADE pairs that were found in neither dataset. Asterisks indicate medication-ADE pairs whose relative rate between the 2 datasets was significantly greater than or less than 1.

than in the claims dataset [relative rate (RR) 2.7, 95% CI, 0.96–7.6]; several other potential ADEs were identified exclusively in the EHR dataset. When the PDC minimum was lowered or potential ADEs occurring during short treatment gaps of between 4 and 25 days were not counted, the relative rate of reflux with sildenafil was significantly greater in the EHR dataset compared to the claims dataset (RR 1.3, 95% CI 1.0–1.6 under both scenarios).

## DISCUSSION

There are notable differences in ascertainment of potential ADE rates across different sources of RWD in children with PH. Potential ADE rates differed by up to 7-fold, and up to 30% of events of interest may be represented only in clinical notes. In general, more

potential ADEs were found in free-text clinical notes than in diagnostic codes data, but certain potential ADEs, such as hearing loss associated with sildenafil exposure, were consistently found more often using diagnostic codes than free-text notes. This analysis, based on noisy RWD with a precision of 69%, may somewhat overestimate the greater representation of potential ADEs in clinical notes, but the overall trend was robust to multiple different analyses. As RWD increasingly complement RCTs in the drug approval pipeline, pharmacoepidemiologists should consider the characteristics of varied data sources for ADE discovery.

Our findings are consistent with prior studies examining the value of different data sources. For instance, Wei and colleagues showed that billing codes and data from clinical notes performed better in combination than either alone for identifying selected diseases in EHRs.[11] Since claims are produced for administration and billing rather than clinical care, they often lack important clinical features and, as such, have limited utility for assessing patient conditions.[23,36–38] Nonetheless, much high-throughput pharmacovigilance continues to be done using claims data alone.[39] Our study quantifies the differences between potential ADE rates ascertained from clinical notes and from administrative claims data and highlights the magnitude of difference in ADE rates that may be found depending on the data source analyzed. Access to structured diagnostic code data in EHRs alone will not suffice to bridge this gap; we found similar patterns of generally more prevalent potential ADEs in clinical notes compared both to EHR diagnostic codes and the external, national claims dataset. Finally, similar patterns of differential ADE representation between structured and unstructured data are evident across different populations.

Though most differentially represented potential ADEs were identified more frequently in clinical notes, several potential ADEs were represented more commonly in diagnostic codes. These differences suggest that clinician behavior in selecting diagnostic codes varies from their practice in creating narrative clinical text. While elucidating specific reasons for this discrepancy requires further study, possible reasons for specific diagnoses being overrepresented in administrative claims data include 1) lack of diagnostic specificity in clinical notes (for example, clinicians do not explicitly label signs/symptoms as a condition, such as "sinusitis," but are steered towards specific diseases when billing), 2) billing for conditions being

ruled out based on signs/symptoms, or 3) billing for reimbursement conditions that are not discussed in the narrative text.

Determining causal relations between medications and signs/symptoms is dependent on clinicians' documented assertions of a sign or symptom being an ADE.[40] Even when the clinician does document, NLP-based methods are not yet always sufficiently robust for RWD-based pharmacovigilance. Even with explicit relations documented, top-performing current NLP systems require extensive customization for the full, integrated task of identifying drug names and explicit mentions of ADEs.[41]

Our study does not attempt to determine a causal relationship between PH-targeted medications and the identified signs and symptoms for either the EHR or claims datasets. A particular advantage of RCTs over RWD is that directly elucidating causal relationships is more straightforward in RCTs. Through randomization and strict reporting controls, an increased rate of adverse events in the treatment group versus the control group should be due to the interventional drug. In contrast, RWD presents challenges, such as the need to tease apart associations from truly causal relations and to avoid confounding. This is often accomplished statistically, indicating an increased frequency of adverse events in patients taking a particular drug, but such studies have at times led to incongruous results. For example, although several clinical trials reported an increased risk of gastrointestinal bleeding with dabigatran versus warfarin, the Mini-Sentinel study found a decreased risk of gastrointestinal bleeding with dabigatran.[42] In the case of PH-targeted drugs in children, the challenge of causal inference with observational data becomes particularly acute, since children with PH who are not prescribed PH-targeted drugs tend to have less severe illness, with fewer comorbidities, than those who are.

Of note, this challenge is relevant for pharmacovigilance studies based on RWD regardless of the data type—structured or unstructured—being used. Nonetheless, pharmacovigilance using RWD is ongoing, whether using diagnostic codes[4] or clinical notes.[1] Although our study does not address the problem of determining whether a documented sign, symptom, or diagnosis represents a true ADE, we do take measures consistent with other studies of NLP-based pharmacovigilance[1] to ensure that the potential ADE is plausible based on temporality and frequency of documentation. The focus of the current work is to compare estimates of potential ADE prevalence depending on the particular RWD type used as a source. Future work should examine the differential ability of different data sources to distinguish true ADEs from unrelated signs and symptoms.

Our study has several additional limitations. The customizations of cTAKES for this specific task used a small number of notes and patients, which risks overtraining of the algorithm. In addition, the NLP pipeline was tuned to maximize sensitivity at the expense of positive predictive value. Some associations between medications and signs/symptoms are thus likely to be spurious, and differences between the EHR clinical notes and diagnostic codes datasets may be exaggerated. Further work is needed to determine gold standard labels for potential ADEs in this population—in particular in RWD for patients with significant comorbidities that may confound causal associations.

## CONCLUSIONS

Analysis of clinical notes generally identifies more potential ADEs than diagnostic codes in either EHR or insurance claims datasets, but certain diagnoses are better represented in structured data. As RWD become a more commonly used data source in pharmacovigilance, researchers and regulators must be cognizant of the high variability in ascertainment rates among different data sources and consider that the strengths and weaknesses of each may vary depending on the diagnosis of interest. The use of multiple data sources may be helpful in making the most accurate measurements. Neither methods using claims data nor processed text can yet adequately distinguish ADEs from disease effects. Further work in determining causality must be addressed in order to expand the use of RWD in pharmacovigilance.

## AUTHOR CONTRIBUTIONS

KDM obtained funding. AG and KDM wrote the manuscript and designed the research. AG, JG, GS, and CL performed the research. AG, SHA, SFM, DDI, MPM, and KDM analyzed the data. GS and CL contributed new analytical tools. SHA, SFM, DDI, MPM, JG, CL, and CKS critically revised the manuscript for important intellectual content. All the authors take responsibility for the final approval of the version to be published and are accountable for all aspects of the work.

## SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

## CONFLICT OF INTEREST STATEMENT

Dr Abman reports receiving laboratory research grants from Shire and United Therapeutics, compensation for DSMB participation from Actelion and United Therapeutics, and an education grant for a Young Investigators Forum from Mallinckrodt. Dr Ivy's institution received fees for Dr Ivy consulting for Actelion, Lilly, and United Therapeutics. Dr Mandl's employer, Boston Children's Hospital, receives corporate philanthropy from Lilly, which makes Adcirca, a brand name of tadalafil. The remaining authors declare that they have no financial or other conflicts of interest.

## REFERENCES

1.  Luo Y, Thompson WK, Herr TM, *et al*. Natural language processing for EHR-based pharmacovigilance: a structured review. *Drug Saf* 2017; 40 (11): 1075–89.
2.  Maxey DM, Ivy DD, Ogawa MT, *et al*. Food and Drug Administration (FDA) postmarket reported side effects and adverse events associated with pulmonary hypertension therapy in pediatric patients. *Pediatr Cardiol* 2013; 34 (7): 1628–36.
3.  Khozin S, Blumenthal GM, Pazdur R. Real-world data for clinical evidence generation in oncology. *J Natl Cancer Inst* 2017; 109 (11). doi: 10.1093/jnci/djx187.

4. Platt R, Brown JS, Robb M, *et al*. The FDA sentinel initiative-an evolving national resource. *N Engl J Med* 2018; 379 (22): 2091–3.

5. Nadkarni PM. Drug safety surveillance using de-identified EMR and claims data: issues and challenges. *J Am Med Inform Assoc* 2010; 17 (6): 671–4.

6. Kim Y, Meystre SM. Ensemble method-based extraction of medication and related information from clinical texts. *J Am Med Inform Assoc* 2019. doi: 10.1093/jamia/ocz100.

7. Cocos A, Fiks AG, Masino AJ. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. *J Am Med Inform Assoc* 2017; 24 (4): 813–21.

8. Tang H, Solti I, Kirkendall E, *et al*. Leveraging Food and Drug Administration adverse event reports for the automated monitoring of electronic health records in a pediatric hospital. *Biomed Inform Insights* 2017; 9: 1178222617713018.

9. Singh JA, Holmgren AR, Noorbaloochi S. Accuracy of Veterans Administration databases for a diagnosis of rheumatoid arthritis. *Arthritis Rheum* 2004; 51 (6): 952–7.

10. Carrell DS, Cronkite D, Palmer RE, *et al*. Using natural language processing to identify problem usage of prescription opioids. *Int J Med Inform* 2015; 84 (12): 1057–64.

11. Wei WQ, Teixeira PL, Mo H, *et al*. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc* 2016; 23 (e1): e20–7.

12. Zeng Z, Deng Y, Li X, *et al*. Natural language processing for EHR-based computational phenotyping. *IEEE/ACM Trans Comput Biol Bioinf* 2019; 16 (1): 139–53.

13. Geva A, Gronsbell JL, Cai T, *et al*. A computable phenotype improves cohort ascertainment in a pediatric pulmonary hypertension registry. *J Pediatr* 2017; 188: 224–31.e5.

14. Klann JG, Abend A, Raghavan VA, *et al*. Data interchange using i2b2. *J Am Med Inform Assoc* 2016; 23 (5): 909–15.

15. Wagholikar KB, Mandel JC, Klann JG, *et al*. SMART-on-FHIR implemented over i2b2. *J Am Med Inform Assoc* 2017; 24 (2): 398–402.

16. Barst RJ, Beghetti M, Pulido T, *et al*. STARTS-2: long-term survival with oral sildenafil monotherapy in treatment-naive pediatric pulmonary arterial hypertension. *Circulation* 2014; 129 (19): 1914–23.

17. Takatsuki S, Calderbank M, Ivy DD. Initial experience with tadalafil in pediatric pulmonary arterial hypertension. *Pediatr Cardiol* 2012; 33 (5): 683–8.

18. Rosenzweig EB, Ivy DD, Widlitz A, *et al*. Effects of long-term bosentan in children with pulmonary arterial hypertension. *J Am Coll Cardiol* 2005; 46 (4): 697–704.

19. Ivy DD, Rosenzweig EB, Lemarie JC, *et al*. Long-term outcomes in children with pulmonary arterial hypertension treated with bosentan in real-world clinical settings. *Am J Cardiol* 2010; 106 (9): 1332–8.

20. Berger RM, Haworth SG, Bonnet D, *et al*. FUTURE-2: Results from an open-label, long-term safety and tolerability extension study using the pediatric FormUlation of bosenTan in pUlmonary arterial hypeRtEnsion. *Int J Cardiol* 2016; 202: 52–8.

21. Takatsuki S, Rosenzweig EB, Zuckerman W, *et al*. Clinical safety, pharmacokinetics, and efficacy of ambrisentan therapy in children with pulmonary arterial hypertension. *Pediatr Pulmonol* 2013; 48 (1): 27–34.

22. Krishnan U, Takatsuki S, Ivy DD, *et al*. Effectiveness and safety of inhaled treprostinil for the treatment of pulmonary arterial hypertension in children. *Am J Cardiol* 2012; 110 (11): 1704–9.

23. Lin C, Dligach D, Miller TA, *et al*. Multilayered temporal modeling for the clinical domain. *J Am Med Inform Assoc* 2016; 23 (2): 387–95.

24. Savova GK, Masanz JJ, Ogren PV, *et al*. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; 17 (5): 507–13.

25. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004; 32 (Database issue): D267–70.

26. Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. *J Biomed Inform* 2003; 36 (6): 414–32.

27. Quan H, Khan N, Hemmelgarn BR, *et al*. Validation of a case definition to define hypertension using administrative data. *Hypertension* 2009; 54 (6): 1423–8.

28. Rector TS, Wickstrom SL, Shah M, *et al*. Specificity and sensitivity of claims-based algorithms for identifying members of Medicare+Choice health plans that have chronic medical conditions. *Health Serv Res* 2004; 39(6 Pt 1): 1839–57.

29. Ong MS, Mullen MP, Austin ED, *et al*. Learning a comorbidity-driven taxonomy of pediatric pulmonary hypertension. *Circ Res* 2017; 121 (4): 341–53.

30. Xie Z, St Clair P, Goldman DP, *et al*. Racial and ethnic disparities in medication adherence among privately insured patients in the United States. *PLoS One* 2019; 14 (2): e0212117.

31. Sattler EL, Lee JS, Perri M 3rd. Medication (re)fill adherence measures derived from pharmacy claims data in older Americans: a review of the literature. *Drugs Aging* 2013; 30 (6): 383–99.

32. Peacock E, Krousel-Wood M. Adherence to antihypertensive therapy. *Med Clin North Am* 2017; 101 (1): 229–45.

33. Frigge M, Hoaglin DC, Iglewicz B. Some implementations of the boxplot. *Am Stat* 1989; 43 (1): 50–4.

34. Koziol N, Bilder C. MRCV: methods for analyzing multiple response categorical variables (MRCVs) [program]. R package version 0.3-3 version; 2014. https://CRAN.R-project.org/package=MRCV.

35. R Core Team. *R: A Language and Environment for Statistical Computing* [*Program*]. Vienna: R Foundation for Statistical Computing; 2018. https://www.-R-project.org.

36. Stein JD, Lum F, Lee PP, *et al*. Use of health care claims data to study patients with ophthalmologic conditions. *Ophthalmology* 2014; 121 (5): 1134–41.

37. Mahmoudi E, Kotsis SV, Chung KC. A review of the use of medicare claims data in plastic surgery outcomes research. *Plast Reconstr Surg Glob Open* 2015; 3 (10): e530.

38. Onukwugha E, Yong C, Hussain A, *et al*. Concordance between administrative claims and registry data for identifying metastasis to the bone: an exploratory analysis in prostate cancer. *BMC Med Res Methodol* 2014; 14 (1)

39. Corrigan-Curay J, Sacks L, Woodcock J. Real-world evidence and real-world data for evaluating drug safety and effectiveness. *JAMA* 2018; 320 (9): 867.

40. Chapman AB, Peterson KS, Alba PR, *et al*. Detecting adverse drug events with rapidly trained classification models. *Drug Saf* 2019; 42 (1): 147.

41. Jagannatha A, Liu F, Liu W, *et al*. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). *Drug Saf* 2019; 42 (1): 99.

42. Sipahi I, Celik S, Tozun N. A comparison of results of the US Food and Drug Administration's Mini-Sentinel Program with randomized clinical trials: the case of gastrointestinal tract bleeding with dabigatran. *JAMA Intern Med* 2014; 174 (1): 150–1.