Integrating Electronic Health Record Data into the ADEpedia-on-OHDSI Platform for Improved Signal Detection: A Case Study of Immune-related Adverse Events

Yue Yu, PhD1, Kathryn J. Ruddy, MD, MPH2, Andrew Wen, MS1, Nansu Zong, PhD1, Shintaro, Tsuji1, PhD, Jun Chen1, PhD, Nilay D. Shah, PhD1,3, Guoqian Jiang, MD, PhD1

1Department of Health Sciences Research, Mayo Clinic, Rochester, MN; 2Department of Oncology, Mayo Clinic, Rochester, MN; 3Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN

Abstract
With widespread adoption of electronic health records (EHRs), Real World Data and Real World Evidence (RWE) have been increasingly used by FDA for evaluating drug safety and effectiveness. However, integration of heterogeneous drug safety data sources and systems remains an impediment for effective pharmacovigilance studies. In an ongoing project, we have developed a next generation pharmacovigilance signal detection framework known as ADEpedia-on-OHDSI using the OMOP common data model (CDM). The objective of the study is to demonstrate the feasibility of the framework for integrating both spontaneous reporting data and EHR data for improved signal detection with a case study of immune-related adverse events. We first loaded the OMOP CDM with both recent and legacy FAERS (FDA Adverse Event Reporting System) data (from the time period between Jan. 2004 and Dec. 2018). We also integrated the clinical data from the Mayo Clinic EHR system for six oncological immunotherapy drugs. We implemented a signal detection algorithm and compared the timelines of positive signals detected from both FAERS and EHR data. We found that the signals detected from EHRs are 4 months earlier than signals detected from FAERS database (depending on the signal detection methods used) for the ipilimumab-induced hypopituitarism. Our CDM-based approach would be useful to provide a scalable solution to integrate both drug safety data and EHR data to generate RWE for improved signal detection.

Introduction
With widespread adoption of electronic health records (EHRs), real world data (RWD) and real world evidence (RWE) have been increasingly used by FDA for evaluating drug safety and effectiveness1. According to FDA2,3, RWD is defined as the data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources including EHRs, whereas RWE is the clinical evidence regarding the usage and potential benefits or risks of a medical product derived from analysis of RWD. FDA has created a RWE program4 to investigate the potential of integrating RWD like EHR data with other drug safety data to support many types of study designs (e.g., randomized pragmatic clinical trials, or observational studies) to generate RWE for regulation decisions.

In fact, it is well recognized in the pharmacovigilance research community that multiple data sources including EHRs are preferred in drug safety surveillance due to inadequacies of single source. For example, Wang, et al.5 performed a case study of signal detection for conventional disease-modifying antirheumatic drugs in rheumatoid arthritis by combining both spontaneous reports and EHR data. Tang, et al.6, leveraged FDA AERS reports for automated monitoring of EHR for adverse drug events (ADEs). Patadia et al.7 evaluated performance of real world EHRs and spontaneous reporting data in drug safety signal detection in the exploring and understanding adverse
drug reactions by integrative mining of clinical records and biomedical knowledge (EU-ADR) project. Iyer, et al.\(^8\) mined clinical text for signals of adverse drug-drug interactions. However, integration of heterogeneous drug safety data sources and systems remains an impediment for effective drug safety surveillance and pharmacovigilance studies. Moreover, social media data is also widely used for ADE signal detection\(^9, 10, 11\). But it is difficult to compare distinct systems and their performances by using various social media data sources\(^9\).

Fortunately, standards-based interoperability solutions based on common data models (CDM) have emerged to tackle the data integration challenges. Notably, the Observational Health Data Sciences and Informatics (OHDSI)\(^12\) has been established as a multi-stakeholder, interdisciplinary collaborative to create open-source solutions based on the Observational Medical Outcome Partnership (OMOP) CDM, which brings out the value of observational health data through large-scale analytics. The OMOP CDM is a deep information model that specifies how to encode and store clinical data at a fine-grained level, ensuring that the same query can be applied consistently to databases around world\(^13\). OHDSI has become an open collaborative with 200 researchers from 25 countries and with 1.26 billion patient records on about 400 million unique patients in its distributed data networks. The OMOP CDM provides a standardized data interface that supports organizing clinical research data into a standard structure in an integrated data repository. The OMOP CDM-based open-source informatics infrastructure and tools have been adopted by a number of large research consortia including the All of Us Research Program\(^14\) (i.e., the Precision Medicine Initiative) and the Electronic Medical Records and Genomics (eMERGE) Research Network\(^15, 16\). This makes large-scale international observational research feasible. Preliminary studies have demonstrated the viability of the CDM-based approaches on the drug safety surveillance studies. For example, Ryan et al.\(^17\) utilized data covering over 130 million cases generated from an amalgamation of the data contained within 10 OMOP CDM-based observational healthcare databases to assess the performance of eight distinct analytic approaches towards adverse event risk identification. Vashishtet al.\(^18\) performed three different studies based on the data from OHDSI community to identify which drug classes among sulfonylureas, dipeptidyl peptidase 4 (DPP-4) inhibitors, and thiazolidinediones are associated with reduced hemoglobin A\(_{1c}\) (HbA\(_{1c}\)) levels and a lower risk of myocardial infarction, kidney disorders, and eye disorders in patients with T2D treated with metformin as a first-line therapy.

In an ongoing project, we have developed a next generation pharmacovigilance signal detection platform known as ADEpedia-on-OHDSI\(^19\) using the OMOP CDM, where open-source extraction, transformation and loading (ETL) tool\(^20\) was developed to convert FAERS data into the OMOP CDM. We aimed to facilitate seamless integration and combined analyses of both Spontaneous Reporting System (SRS) data such as FDA Adverse Event Reporting System (FAERS) and EHR data for modern pharmacovigilance studies because of the following three reasons: 1) Some data in the FAERS may not be standardized, such as drug name, which could be the brand name, ingredient name, or even spelling error. That will bring bias for the data collection and analysis. So, it is important to standardize the FAERS data into a standard driven CDM such as OHDSI CDM. 2) By integrating FAERS data and EHR data together into one database, that would facilitate integrative data mining. 3) Using the CDM-based data platform, we could design standard data retrieval queries, which could be reused in future studies, and the queries also could be shared with the research communities. The details regarding the FAERS ETL processing and validation could be found in our previously published paper\(^19\).

The objective of the study is to demonstrate the feasibility of the framework for integrating both FAERS data and EHR data for improved signal detection with a case study of immune-related adverse events (irAEs). Specifically, we focused on a FAERS confirmed irAE, ipilimumab-induced hypopituitarism, as our study target in this research. We first collected normalized FAERS data and implemented a classical FAERS based ADE signal detection algorithm, Reporting Odds Ratio (ROR), to detect the ipilimumab-induced hypopituitarism. To show the signal detection ability of the normalized EHR data in our platform, we also conducted a case-control study. We compare the timeline of the irAE signal detection between normalized FAERS data and EHR data to show the capacity of signal detection in different data sources. This paper is a preliminary research to show the feasibility of our platform in the next-generation ADE signal detection.

Materials and Methods

Materials

FAERS

FAERS\(^21\) is a database that contains adverse event reports, medication error reports, and product quality complaints resulting in adverse events that were submitted to FDA. FAERS provides rich information on voluntary reports of
suspected adverse events and has been widely used for drug safety signal detection and pharmacovigilance applications. ADEs could be spontaneously reported by drug manufacturers, physicians, pharmacists, nurses, and consumers. There are seven tables in FAERS and they will update quarterly. The main limitation of the FAERS database is that it contains non-standardization of the data, lacks of a baseline number of patients taking the medication, with the potential of the missing reports. To overcome those limitations, we propose to develop a standard driven platform to perform data standardization and comprehensive ADE signal detection using both FAERS data and EHR data. In this study, we used the FAERS data covering the time period from January 2004 to December 2018.

**OMOP CDM**

The OMOP CDM is a data model which focuses on transforming different observational databases into a common format and a common representation to allows for the systematic analysis of disparate observational databases. The CDM defines table schemas in a person-centric manner. As of Oct. 11, 2018, version V6.0 of the CDM was released, containing 39 tables in 6 categories: standardized clinical data, standardized health system data, standardized health economics, standardized metadata, standardized vocabularies and standardized derived elements. Notably, terminology normalization enabled by standard vocabularies with focus on SNOMED CT, LOINC, and RxNorm is a strong characteristic of the OMOP CDM.

**Mayo Clinic EHRs**

Mayo Clinic is an academic medical center that provides comprehensive patient care, education in clinical medicine and medical sciences, and extensive programs in research. The Mayo Clinic Unified Data Platform (UDP) has been implemented to provide practical data solutions and creates a combined view of multiple heterogeneous EHR data sources (including Epic) through effective data orchestration, along with a number of data marts based on common data models. The UDP serves as a data warehouse that contains millions of patients’ data for the support of both clinical practice and research. UDP is also updated in real time.

**Methods**

**Data integration in the ADEpedia-on-OHDSI platform**

In this study, the SRS data and the EHR data were standardized and integrated into the ADEpedia-on-OHDSI platform. Figure 1 illustrates the integration process of the ADEpedia-on-OHDSI platform. First, we used the ETL tool developed for the ADEpedia-on-OHDSI platform to convert FAERS data from January 2004 to December 2018 into OMOP CDM. Then, for EHR data, we collected the data of the patients who were administered with any of the six FDA approved immunotherapy drugs (i.e., ipilimumab, pembrolizumab, nivolumab, atezolizumab, durvalumab, and avelumab) from the Mayo Clinic UDP. We also implemented an ETL process to transform those EHR data into the OMOP CDM. After the ETL process, both the FAERS and EHR data were loaded into the ADEpedia-on-OHDSI platform that uses the same data model and vocabulary standards.

![Figure 1. ADEpedia-on-OHDSI framework](image-url)
Data Analysis

We focused on an irAE known as ipilimumab-induced hypopituitarism as our study target in this research. We chose ipilimumab because it was the first oncological immunotherapy drug approved by the US FDA in 2011 to treat melanoma. Therefore, in comparison with other immunotherapy drugs, there are enough ipilimumab administration cases for data analysis. Hypopituitarism is one of the most common irAEs that induced by ipilimumab, and it was reported by both FDA drug label\textsuperscript{22} and previous publications\textsuperscript{23}. In this study, we implemented different signal detection algorithms and compared the timelines of positive signals detected using standardized data in the ADEpedia-on-OHDSI platform. The goal is to evaluate the effectiveness of adverse event signal detection in the different data sources and different signal detection methods. Figure 2 shows the study design of our irAEs signal detection study.

Figure 2. Study design for ipilimumab-induced hypopituitarism signal detection

First, we designed a case-control group comparison using EHR data. The case-control comparison design is based on the contingency table as shown in Table 1. For the case group, we collected all the patients’ data who had ipilimumab for their treatment. Patients who had at least one drug administration but did not take ipilimumab as a treatment were selected as controls. Then, hypopituitarism after patients taking an ipilimumab administration/other administration for melanoma were seen as an exposure condition. Odds ratio between the case and control groups were calculated separately for standardized FAERS data and EHR data. Fisher’s test was conducted to assess if there was a significant difference for patients with hypopituitarism between case/control group. We also performed Bonferroni correction to adjust the p-value. When the lower 95\% CI of OR is greater than 1 and the adjusted p-value is less than 0.05, we considered a positive adverse event signal detected. We used cumulative data to calculate the OR and p-value monthly from Jan. 2011 and generated a signal detection timeline to observe the first date of a positive signal detected.

Table 1. Contingency table of the Case-Control comparison.

<table>
<thead>
<tr>
<th></th>
<th>Cases: Patients with ipilimumab administration</th>
<th>Controls: Patients without ipilimumab administration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients with hypopituitarism after administration</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Patients without hypopituitarism after administration</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>
Furthermore, due to the fact that FAERS only recorded the report of patients who suffered an adverse event, the baseline number of patients who used the target drug is not available in FAERS, which may cause a bias in detecting adverse event signals by conducting a case-control comparison using FAERS data. So, we also implemented a widely used adverse event signal detection method, Reporting Odds Ratio (ROR), to evaluate the signal detection reaction time and effect size. The ROR is calculated by the following contingency table (Table 2) and Equation (Equation 1). When the target drug-event pair’s lower limit of 95% CI of ROR > 1, we considered a positive signal detected. Similarly, for the case-control comparison, we also used the cumulative data to calculate the ROR monthly from Jan. 2011 to Dec. 2018 to evaluate the ability of ROR in the adverse event signal detection.

Table 2. Contingency table for the ROR calculation.

<table>
<thead>
<tr>
<th>Patient’s reports with Hypopituitarism</th>
<th>Patient’s reports without ipilimumab</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a</td>
</tr>
<tr>
<td>Patient’s reports without Hypopituitarism</td>
<td>c</td>
</tr>
</tbody>
</table>

\[
ROR = \frac{a/b}{c/d} \quad \text{(Equation 1)}
\]

Results

Data ETL result

For the FAERS database, a total of 9,956,309 patients’ data between Jan. 2004 and Dec. 2018 were successfully standardized and transformed into the ADEpedia-on-OHDSI platform. In addition, those patients were associated with 37,288,989 records of drug exposures and 32,504,326 records of adverse events. All of the standard FAERS data were loaded into 8 different OMOP tables. The details of the ETL process evaluation could be found in our previous study[16].

For EHR data, we identified 761 patients who had taken six oncological immunotherapy drugs (i.e., ipilimumab, pembrolizumab, nivolumab, atezolizumab, durvalumab, and avelumab) from the Mayo Clinic UDP. The first dates of immunotherapy drug administration for those patients were between Mar. 18th 2010 and Feb. 5th 2019. We also collected and transformed 443,647 records of diagnosis data and 373,074 records of drug administration data from those 761 patients into the ADEpedia-on-OHDSI platform.

Description of the data processing

Figure 3 shows the data collection results for the case-control comparison. We analyzed the distributions of both the case group data and the control group data as we mentioned in the methods section. Those 4 bar plots in Figure 3 were used to illustrate the distributions of data in the four cells of the contingency table (Table 1 and Table 2) which we employed to detect the ipilimumab-induced hypopituitarism signal. The cumulative case numbers of each group were shown yearly from Jan. 2011 to Dec. 2018. The red bar and the blue bar represented that the data were collected from EHR and FAERS respectively. Figure 3a and Figure 3c displayed the number of cases of the patients who took the ipilimumab. These figures indicated that there were a significantly greater number of cases in the FAERS than the EHR data. This is due to that the FAERS is a global database which gathers the adverse event reports from all over the world, whereas our EHR data is only collected from several medical centers of the Mayo Clinic. The difference in detected cases without ipilimumab administration between the EHR and FAERS as indicated in Figure 3d can also be explained by the same reason. Besides, the earliest date of the cases who took ipilimumab as a treatment in the FAERS was Nov. 2005, whereas the earliest date of the ipilimumab administration in our EHR was Mar. 2010. This is because FAERS also collected the adverse reports of the clinical trials from worldwide. On the contrary, the Figure 3b showed that the number of patients who had hypopituitarism but did not take ipilimumab is significantly more in the EHR than that in the FAERS, which was caused by the reason that the FAERS only recorded the adverse event reports in specific patients but couldn’t cover all the patients who took the
corresponding medications. The EHR, by contrast, could provide a more accurate case number of drug administrations as a study baseline.

**Figure 3.** Cumulative case number of case-control comparison. Figure 3a indicates the patients in the case group which meet the exposure condition; 3b indicates the patients in the control group which meet the exposure condition; 3c indicates the patients in the case group which do not meet the exposure condition; and 3d indicates the patients in the control group which do not meet the exposure condition. The red bar represents the cases and controls from EHR; The blue bar represents the FAERS cases and controls.

**Ipilimumab-induced Hypopituitarism Signal Detection Results**

Figure 4 demonstrates the ipilimumab-induced Hypopituitarism signal detection results for different analytical methods and different data source. According to the positive signal detection criteria we defined in the methods section, the earliest date of the positive signal we detected is Mar. 2011 by the case-control comparison of EHR data, 4 months earlier than that using FAERS and ROR for the detection. For the effect size value (OR or ROR), in most cases, the ROR value is greater than the OR value calculated by a case-control comparison.
Figure 4. The timelines of the ipilimumab-induced Hypopituitarism signal detection results. The dot and the line show the Log OR (or ROR) and the 95% CI value. The red line displays the detection results of the case-control comparison by CDM-based EHR data, the red date shows the first detection date of positive signal by EHR; The blue line displays the results of ROR-based detection by CDM-based FAERS data, the blue date shows the first detection date of positive signal by FAERS.

Abbreviation: OR, Odds Ratio; ROR, Reporting Odds Ratio; EHR, Electronic Health Record; FAERS, FDA Adverse Event Reporting System; 95% CI, 95 Confidence Interval.

Discussion

In this study, we utilized our standards-driven pharmacovigilance platform ADEpedia-on-OHDSI to implement a signal detection algorithm and compared the timelines of positive signals detected from both FAERS and EHR data. We also evaluated the effectiveness of different detection methods and data sources. Previous studies demonstrated that specific adverse event signals would be detected earlier if using the EHR data rather than using the SRS data. For example, rofecoxib, a non-steroidal anti-inflammatory drug, was withdrawn in 2004, 5 years after it was approved by FDA, due to its severe cardiovascular adverse events. By the time it was withdrawn, it was estimated that 80 million people had taken rofecoxib\(^2\). However, some studies had noted that with a real-world data on 100 million patients, a statistically significant “signal” of serious cardiovascular risk could have been detected after less than 3 months of exposure with rofecoxib\(^2\). Patadia et al.\(^2\) further proved the point by implementing a signal detection approach by the use of both SRS and EHR data. They used EU-ADR database (for EHR) and WHO-VigiBase (for SRS) to detect the association between Rofecoxib and acute myocardial infarction. And they found that the use of EU-EHR databases was able to detect the AMI signal 4 years earlier than the use of the SRS databases. We argue that there are still very few studies regarding the discovery with respect to differences in early signal detection between EHR and SRS and that there exists a need to conduct further comparative studies for differing adverse events, particularly for those adverse events induced by newly approved drugs. According to our signal detection results, the case-control comparison using EHR data showed the fastest reaction time for signal detection. Utilizing SRS for the signal detection (one of the most important drug post-marketing pharmacovigilance methods), we found that there is a 4 more months delay for detecting the same positive signal in FAERS in comparison with the signal detected using the EHR data. This delay was probably because it would cost several months for FDA to capture the adverse event reports. These findings provide clear evidence demonstrating the importance of the EHR data on early signal detection of irAEs for the post-market drug safety surveillance studies.
For the effectiveness evaluation, the EHR-based case-control comparison indicated the most accurate effect size for the signal detection. Since the number of cases without target drug and target adverse event (the value in the column “d” in Table 2) was tremendously large, the effect size of ROR calculated using the FAERS data was much higher than that using the EHR data. Although our research finding indicated that EHR-based signal detection would have a better signal detection reaction time and effectiveness than the SRS-based approach, a smaller number of adverse event reports made it more unfavorable to detect those rare adverse events. Therefore, signal detection using both SRS and EHR would strengthen current signal detection activities by decreasing the influences of systematic bias when using one database.

However, one of the biggest challenges for the heterogeneous database integration is data standardization. For example, in this study, the drug names in raw FAERS are free text, while the drug names in our EHR are coded by RxNorm. So, through converting those databases into the OMOP CDM, the medical concepts can be standardized using the same set of standard vocabularies. In addition, data standardization using the CDM-based approach also provides a mechanism that we could collect and integrate more EHR drug safety data from multiple institutions. At the same time, the standard queries and applications could also be reused across institutions and research communities, which may save a lot of time and resources for researchers in collaborations.

In this study, one of the limitations we met was regarding the storage of the FAERS’ adverse event report dates. During the ETL process of the FAERS, we loaded all the adverse event data into the “OBSERVATION” table of the OMOP CDM. And there is only one field, observation_date, for recording the adverse event occurrence dates. In this research, we also need the date when FDA received adverse event cases to evaluate the first positive signal detection time. In order to solve this problem, we created a workaround, i.e., storing the FDA report receiving date into the “value_as_datetime” field. In future work, we will work with the OHDSI community to come up with a better solution such as adding a new field in the OBSERVATION table to solve the problem. Another limitation is about our case-control design, we should consider some confounding factors that could bias the analysis for the cases and the controls. Those confounding factors include patients’ age, gender, race, social determinants, and healthcare history. Other confounding factors such as dose response and the temporality of the case-control design would also be considered in our future analysis.

Our research validated the feasibility of detecting adverse event signals by integrating both SRS and EHR data using OMOP CDM. Our approach could be generalized easily to study other irAEs in a scalable manner. Both our FAERS data transformation tool and the OMOP CDM are open source and publicly available. Researchers could use these open source tools to integrate their own datasets for similar studies. We will collaborate with both the pharmacovigilance research and OHDSI communities in the future to improve the scalability of our ADEpedia-on-OHDSI platform for effective drug safety studies. In addition, we will design some more comprehensive ADE signal detection methods such as cohort or time-series analyses for our ADEpedia-on-OHDSI platform in the future work.

Conclusion

In this study, we have demonstrated the feasibility of the ADEpedia-on-OHDSI framework for integrating both spontaneous reporting data and EHR data for improved signal detection with a case study of immune-related adverse events. We implemented a signal detection algorithm and compared the timelines and effectiveness of positive signals detected from both FAERS and EHR data. The outcome of this research indicated that our CDM-based approach would be useful to provide a scalable solution to integrate both drug safety data and EHR data to generate RWE for improved signal detection.

Acknowledgement

This research was funded in part by FDA HHSF223201710167C and NIH BD2K U01 HG009450.

References

Health Records Databases Complement Spontaneous Reporting System Databases? A Historical
Med. 2007;356(17):1700

McClellan M. Drug safety reform at the FDA
rofecoxib: cumulative meta
Juni P, Nartey L, Reichenbach S, Sterchi R, Dieppe PA, Egger M. Risk of cardiovascular events and

2019;103119.

Yu, Y. and G. Jiang. ADEpedia-on-OHDSI. Available from:
https://github.com/adepedia/adepedia-on-ohdsi.

FDA. FDA Adverse Event Reporting System (FAERS). Available from:
default.htm.

NIH. Drug Label: YERVOY- ipilimumab injection. Available from:

Araujo PB, Coelho MC, Arruda M, Gadilha MR, Neto LV. Ipilimumab-induced hypophysitis: review of

Juni P, Narrey L, Reichenbach S, Sterchi R, Dieppe PA, Egger M. Risk of cardiovascular events and

McClellan M. Drug safety reform at the FDA - Pendulum swing of systematic improvement? New Engl J

Health Records Databases Complement Spontaneous Reporting System Databases? A Historical-

718