Article published online: 2025-02-12

Discrepancies in Aggregate Patient Data between Two Sources with Data Originating from the Same Electronic Health Record: A Case Study

Allen J. Yiu[1,2,3] Graham Stephenson[1] Emilie Chow[4] Ryan O'Connell[1,5]

[1] Department of Emergency Medicine, University of California, Irvine, California, United States
[2] Department of Pediatrics, Children's National Hospital, Washington, District of Columbia, United States
[3] Department of Pediatrics, George Washington University School of Medicine and Health Sciences, Washington, District of Columbia, United States
[4] Department of Medicine, University of California, Irvine, California, United States
[5] Department of Pathology, University of California, Irvine, California, United States

Address for correspondence Allen J. Yiu, MD, MBA, FAAP, Department of PediatricsChildren's National Hospital, George Washington University School of Medicine and Health Sciences, 111 Michigan Avenue NW, Washington, DC 20010, United States (e-mail: ayiu@childrensnational.org).

Appl Clin Inform 2025;16:137–144.

Abstract

Background Data exploration in modern electronic health records (EHRs) is often aided by user-friendly graphical interfaces providing "self-service" tools for end users to extract data for quality improvement, patient safety, and research without prerequisite training in database querying. Other resources within the same institution, such as Honest Brokers, may extract data sourced from the same EHR but obtain different results leading to questions of data completeness and correctness.

Objectives Our objectives were to (1) examine the differences in aggregate output generated by a "self-service" graphical interface data extraction tool and our institution's clinical data warehouse (CDW), sourced from the same database, and (2) examine the causative factors that may have contributed to these differences.

Methods Aggregate demographic data of patients who received influenza vaccines at three static clinics and three drive-through clinics in similar locations between August 2020 and December 2020 was extracted separately from our institution's EHR data exploration tool and our CDW by our organization's Honest Brokers System. We reviewed the aggregate outputs, sliced by demographics and vaccination sites, to determine potential differences between the two outputs. We examined the underlying data model, identifying the source of each database.

Results We observed discrepancies in patient volumes between the two sources, with variations in demographic information, such as age, race, ethnicity, and primary language. These variations could potentially influence research outcomes and interpretations.

Keywords
- electronic health record
- databases
- data quality
- data completeness
- data analysis

received
July 14, 2024
accepted after revision
September 4, 2024

DOI https://doi.org/10.1055/a-2441-3677.
ISSN 1869-0327.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Conclusion This case study underscores the need for a thorough examination of data quality and the implementation of comprehensive user education to ensure accurate data extraction and interpretation. Enhancing data standardization and validation processes is crucial for supporting reliable research and informed decision-making, particularly if demographic data may be used to support targeted efforts for a specific population in research or quality improvement initiatives.

Background and Significance

The adoption of electronic health records (EHRs) has allowed for the storage and retrieval of vast amounts of patient clinical and health data. Modern EHRs often provide "self-service" data exploration tools enabling users to extract these data for quality improvement, data intelligence, management decision-making, and research.[1,2] While the use of graphical user interface data mining tools has become more widespread, it remains unclear whether the data obtained from these built-in tools align with data sourced through other methods within the same EHR system.

Recent studies have utilized SlicerDicer, Epic's built-in clinical data exploration and analysis tool,[3–5] as their primary source. Within our organization, while SlicerDicer is used to generate hypotheses and conduct feasibility analyses, obtaining patient data for research requires the services of an Honest Broker. In our case, after we utilized SlicerDicer to extract preliminary data, our team identified differences in the aggregate numbers reported by Honest Broker, despite both datasets being sourced from the same EHR. Such discrepancies may not only lead to questions related to data completeness and correctness as defined by prior research[6–9] but may also have broader implications related to research workflow and data utilization in research reporting.

Objectives

Our objectives were to analyze the data discordance between the de-identified aggregate patient information obtained via SlicerDicer and that delivered by an Honest Broker analyst via a query to our clinical data warehouse (CDW), and subsequently determine the potential cause(s) of such differences. By studying the causes, we can ensure greater confidence in the data output and its use.

Methods

Case Study

This case study examines the aggregate-level demographic characteristics of patients who obtained influenza vaccinations at three static and three drive-through clinics between August 2020 and December 2020. Epic's SlicerDicer was used to examine de-identified aggregate patient characteristics who met the criteria for receiving an influenza vaccine at the selected locations within the desired date range. Following approval by our Institutional Review Board (IRB) for a separate research protocol, we sent a request to our institution's Honest Broker to obtain patient-level data with the same criteria. We utilized aggregate patient counts for this study. This study did not constitute human subject research and met the criteria for non-human subjects research (NHSR) self-determination at our institution.

Data Extraction from SlicerDicer

The denominator was chosen for "all patients" between August 2020 and December 2020. We identified specific "Location" that administered static vaccinations and "Department" that administered drive-through vaccinations. Because "Department" are grouped into the higher level "Location," to obtain accurate static site data, drive-through vaccination "Department" were excluded. We refined our search by only including patients whose influenza vaccination was recorded under the category of "Medication," "Procedure," or "Immunization."

Data Extraction from Clinical Data Warehouse

An Honest Broker analyst extracted data from our organization's CDW using a structured query language (SQL) targeting the same criteria. Our organization's CDW utilizes the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) format, standardizing structure and data obtained from multiple heterogenous sources.

Data Model Review

We reviewed the current existing data model in our organization (►Fig. 1). The data for SlicerDicer are sourced through Epic's Caboodle data warehouse, which is created from the EHR's normalized database, Clarity, after an extract transform load (ETL) process is performed per Epic's underlying business rules. Data for OMOP are sourced from the Clarity database, via OMOP's ETL business rules, as well as external billing/claims data, legacy EHR data, and other third-party application data.

Analysis

For this study, we utilized de-identified aggregate data from OMOP to compare the de-identified aggregate data obtained from SlicerDicer. Because we are comparing two values between corresponding variables, we calculated the percentage difference between aggregate results with a threshold of 2% or less as an indicator of data quality, according to published methods.[10]

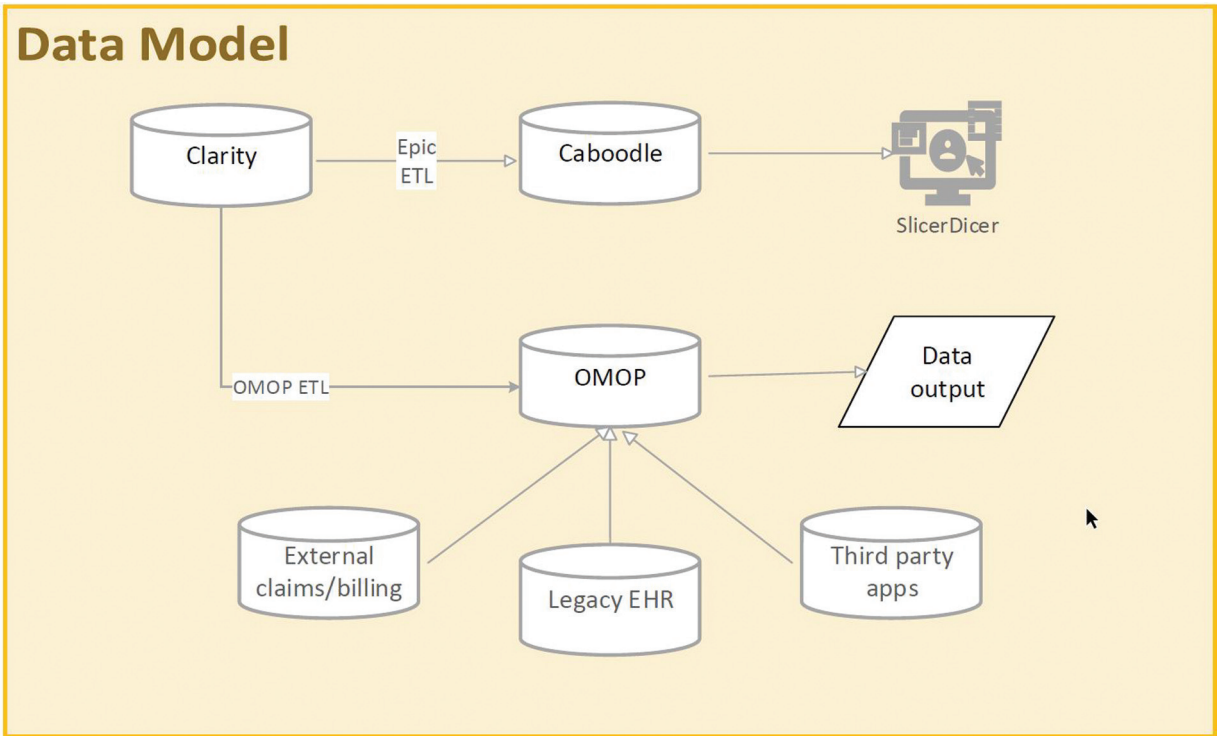

Fig. 1 Data model of OMOP and SlicerDicer showing a shared source for patient data with differing ETL rules. EHR, electronic health record; ETL, extract transform load; OMOP, Observational Medical Outcomes Partnership.

Results

SlicerDicer and Observational Medical Outcomes Partnership Comparison

►**Table 1** shows the difference in aggregate numbers. Broken down by the site of immunization, we found that the difference between SlicerDicer and OMOP is 6.15% and 1.38% when comparing static versus drive-through sites for immunizations, respectively.

In ►**Table 2**, as we further categorized the aggregate data by age, race, ethnicity, sex, and primary language spoken, we saw continued dichotomization. The difference between the two outputs ranged from 0 to 167.57%, with the greatest differences under "Unknown" or "Other" for each demographic category, and for each age category.

When sliced by the site of immunization in ►**Table 3**, the difference in aggregated extracted data between SlicerDicer and OMOP ranged from 0 to 200%. Under race, ethnicity, and primary language spoken, the largest differences between SlicerDicer and OMOP were found in the "Unknown" or "Other" categories. Other notable differences included race, under the "Native Hawaiian/Other Pacific Islander" and "American Indian/Alaska Native" categories; age, generally across all categories; and ethnicity, under the "Non-Hispanic/Latino" category. When sliced by the site of immunization, most differences between the two outputs fell under the static site.

Data Query Review and Comparison

We reviewed the query in SlicerDicer, and the SQL code used to extract data from OMOP. In SlicerDicer, drive-through sites were listed under "Department," and static sites were listed under "Location." "Department" sites were grouped into a higher level "Location" (►**Fig. 2**). Accordingly, utilizing custom logic in SlicerDicer, "Department" drive-through sites were excluded from "Location" static sites when obtaining static site data. Because OMOP only contained "Department" in the data warehouse, individual site IDs were pulled manually from Clarity, and static sites were manually mapped to ensure each static site was appropriately included.

In SlicerDicer, influenza vaccinations were recorded under "Medication," "Procedure," or "Immunization." To

Table 1 Aggregate numbers between SlicerDicer and Observational Medical Outcomes Partnership sliced by immunization site

Immunization site	Source		Percentage difference
	SlicerDicer *N* (%)	OMOP *N* (%)	
Static	10,238 (78.3)	10,888 (79.5)	6.15%
Drive-through	2,839 (21.7)	2,800 (20.5)	1.38%

Abbreviation: OMOP, Observational Medical Outcomes Partnership.

Table 2 Demographic differences in aggregate data between SlicerDicer and Observational Medical Outcomes Partnership

Characteristics	Source		Percentage difference
	SlicerDicer *N* (%)	OMOP *N* (%)	
Age (y)			
≤21	2,520 (19.3)	2,809 (20.5)	10.85%
22–64	5,916 (45.2)	6,777 (49.5)	13.57%
≥65	4,641 (35.5)	4,102 (30.0)	12.33%
Race			
White	8,376 (62.3)	8,377 (61.2)	0.01%
Asian	2,638 (19.6)	2,800 (20.5)	5.96%
Black/African American	315 (2.3)	293 (2.1)	7.24%
Unknown	135 (1.0)	180 (1.3)	28.57%
Native Hawaiian/Other Pacific Islander	87 (0.6)	79 (0.6)	9.64%
American Indian/Alaska Native	52 (0.4)	44 (0.3)	16.67%
Other	1,850 (13.8)	1,915 (14.0)	3.45%
Ethnicity			
Hispanic/Latino	4,783 (36.6)	4,525 (33.1)	5.54%
Non-Hispanic/Latino	8,203 (62.7)	9,046 (66.1)	9.77%
Unknown/other	91 (0.7)	117 (0.8)	25.00%
Sex			
Female	7,330 (56.1)	7,575 (55.4)	3.29%
Male	5,746 (43.9)	6,095 (44.6)	5.89%
Primary language spoken			
English	10,175 (77.8)	11,001 (80.4)	7.80%
Spanish	2,404 (18.4)	2,204 (16.1)	8.68%
Other	495 (3.8)	449 (3.3)	9.75%
Unknown/not recorded	3 (0.0)	34 (0.2)	167.57%

Abbreviation: OMOP, Observational Medical Outcomes Partnership.

produce accurate results, each vaccination site was "linked" to each vaccination listed under "Medication," "Procedure," or "Immunization." The data from Clarity's immunization tables were included in OMOP's procedures and medications tables. OMOP did not have an immunization table.

Additionally, in SlicerDicer, the age shown was the current age of the patient, unless the checkbox "Specify Age at Time of Event" was checked. After the data were pulled from OMOP, age at the time of vaccine administration was manually calculated based on the date of vaccination and date of

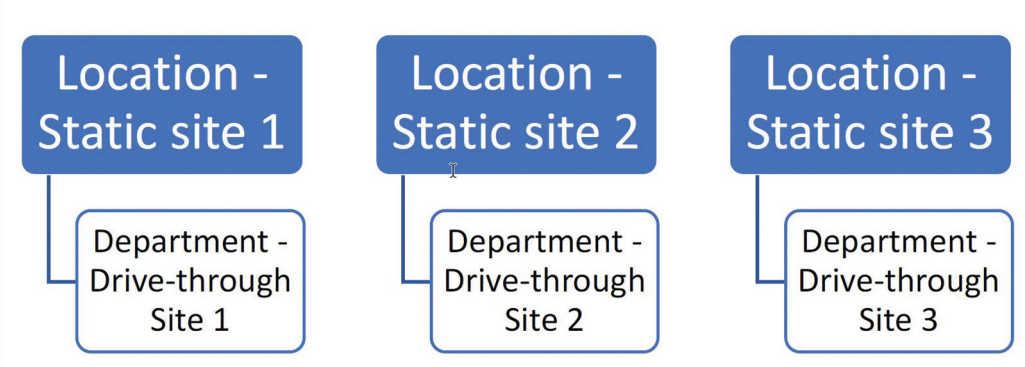

Fig. 2 In SlicerDicer, "Department" sites are grouped into a higher level "Location." The drive-through vaccination sites were listed under "Department" and the static immunization sites were listed under "Location."

Table 3 Demographic differences in aggregate data between SlicerDicer and Observational Medical Outcomes Partnership when sliced by immunization site

Characteristics	Static immunization site			Drive-through immunization site	
	Source			Source	
	SlicerDicer *N* (%)	OMOP *N* (%)	SlicerDicer *N* (%)	OMOP *N* (%)	Percentage difference
Age (y)					
≤21	2,311 (22.5)	2,572 (23.6)	209 (7.5)	237 (8.5)	12.56%
22–64	4,280 (41.6)	5,002 (46.0)	1,636 (58.6)	1,775 (63.4)	8.15%
≥65	3,692 (35.9)	3,314 (30.4)	949 (34.0)	788 (28.1)	18.54%
Race					
White	6,628 (62.5)	6,660 (61.2)	1,748 (61.3)	1,717 (61.3)	1.79%
Asian	1,920 (18.1)	2,104 (19.3)	718 (25.2)	696 (24.9)	3.11%
Black/African American	259 (2.4)	242 (2.2)	56 (2.0)	51 (1.8)	9.35%
Unknown	101 (1.0)	146 (1.3)	34 (1.2)	34 (1.2)	0.00%
Native Hawaiian/Other Pacific Islander	63 (0.6)	60 (0.6)	24 (0.8)	19 (0.7)	23.26%
American Indian/Alaska Native	43 (0.4)	37 (0.3)	9 (0.3)	7 (0.3)	25.00%
Other	1,587 (15.0)	1,639 (15.1)	263 (9.2)	276 (9.9)	4.82%
Ethnicity					
Hispanic/Latino	4,280 (41.6)	4,016 (36.9)	503 (18.0)	509 (18.2)	1.19%
Non-Hispanic/Latino	5,938 (57.7)	6,780 (62.3)	2,265 (81.1)	2,266 (80.9)	0.04%
Unknown/other	65 (0.6)	92 (0.8)	26 (0.9)	25 (0.9)	3.92%
Sex					
Female	5,749 (55.9)	5,991 (55.1)	1,581 (56.6)	1,584 (56.6)	0.19%
Male	4,533 (44.1)	4,879 (44.9)	1,213 (43.4)	1,216 (43.4)	0.25%
Primary language spoken					
English	7,587 (73.8)	8,413 (77.3)	2,588 (92.6)	2,588 (77.3)	0.00%
Spanish	2,258 (22.0)	2,053 (18.9)	146 (5.2)	151 (5.4)	3.37%
Other	435 (4.2)	391 (3.6)	60 (2.1)	58 (2.1)	3.39%
Unknown/not recorded	3 (0.0)	31 (0.3)	0 (0)	3 (0.1)	200.00%

Abbreviation: OMOP, Observational Medical Outcomes Partnership.

birth. Other differences found during the query comparison can be found in ►**Table 4**.

Influenza vaccination was recorded in SlicerDicer and OMOP only after the vaccination was administered. External claims data were excluded from the OMOP query.

Discussion

While data discrepancies have been identified between an organization's EHR and external sources, such as pump infusion logs[11] or multiple cancer registries,[12] and between EHRs across health care systems in shared research networks,[13] our case study highlights discrepancies between two platforms that draw data from a single EHR. These inconsistencies, particularly within certain patient strata, could potentially impact research outcomes and interpretations, raising concerns about data completeness and correctness.

Broadly, such differences may be due to a variety of factors. The secondary use of EHR data for research has inherent limitations, notably related to data quality and preprocessing.[14,15] Data quality and integrity may be affected by unstandardized data collection and documentation practices (patient-specified vs. provider-assumed), separate and disparate purposes for data collection (i.e., patient care vs. financial implications), changes to data collection procedures, improperly matched data elements, and/or variability in data vocabulary and definitions.[16–18]

In our specific case, there was variability in race, ethnicity, and primary language spoken. Inaccurate and discordant race and ethnicity reporting between EHR and self-reporting have been found due to inconsistencies in documentation and recording, and fluid definitions of race and ethnicity,[19–22] particularly with classifications of American Indian and Alaska Native,[23] with greater variability in the race and ethnicity data within EHR data compared with other

Table 4 Summary of data query review between SlicerDicer and Observational Medical Outcomes Partnership

	SlicerDicer	OMOP
Immunization sites	• "Department" sites are grouped into a higher level "Location" • Drive-through vaccination sites listed under "Department." Static immunization sites are listed under "Location" • To obtain each "Location," static immunization site data required "Department" drive-through influenza sites to be excluded from the "Location" static immunization sites utilizing custom logic	• Data table "Department" in the data warehouse • Individual site IDs were pulled manually from Clarity to ensure each "Location" static immunization site was appropriately included in the query
Vaccination	• Recorded under "Medication," "Procedure," or "Immunization" tables • To produce accurate results, each vaccination site was "linked" to influenza vaccination listed under "Medication," "Procedure," or "Immunization"	• Clarity's immunization tables were included in OMOP's procedures and medications tables
Age	• Age shown is the current age of the patient unless the checkbox "Specify Age at Time of Event" was checked	• Age shown is the current age of the patient at the time of data extraction • Calculated based on the date of vaccination and date of birth

Abbreviation: OMOP, Observational Medical Outcomes Partnership.

sources.[12] Additionally, data quality and migration studies have found that imperfect harmonization, misclassification, and non-conforming data with poor categorization and coding schemes may contribute to variances.[24,25] CDM implementation is meant to improve data quality and increase efficiency when performing research, but it may also lead to information loss if data sources do not map to vocabulary concepts or if the rules in the ETL process are inappropriate.[26] Thus, while inconsistencies in data collection and categorization may have affected output, the variability may have also been caused by the data migration process, reaffirming the importance of regular validation of the data and review of ETL processes to ensure data quality and minimize biases.

Further review of race, ethnicity, and primary language spoken shows that the "unknown" or "other" categories had the greatest variance. The use of "unknown" or "other" categorization at the point of data collection may be due to guessing or lack of granulation in the categories available.[12] In our case, when the data were extracted from OMOP, the categories for race, ethnicity, or primary language spoken did not correlate completely with those in the EHR. As a result, the output was often categorized as "other" or "no matching concept." Timely mapping and further granulation of race, ethnicity, and language categories would enhance data quality and integrity.

A granular comparison of race, ethnicity, and primary language spoken also showed variation between static versus drive-through sites. In our organization, while OMOP has a "Department," it does not have a "Location" concept, requiring manual mapping of each static site from Clarity. This manual mapping may explain this variance and lead to false positive or false negative errors,[27] although some studies have shown no difference between automated or manual mapping.[28] Interestingly, there was greater consistency in drive-through sites since the "Department" concept is present in both Epic and OMOP, which may validate the automated transfer of data when concepts are present in both databases.

Lessons Learned

While this report draws on our institution's research workflow as a case study to compare extracted output between a self-service exploration tool and our organization's CDW, the comparison is not meant to extol one platform over another. Indeed, variability in data quality has been found between self-service reporting tools[10] and EHR across health care systems utilizing a single CDM,[13,29] which does not preclude its use. Rather, highlighting the discrepancies should reiterate the importance of user education, examine possible improvements to ETL processes, and prompt further discussions to improve data quality.

User training on self-service tools is necessary to realize their capabilities[30] and ensure accurate output. For example, without the knowledge that drive-through "Department" sites group into "Location," static site output may contain "duplicate" data if drive-through "Department" sites were not properly excluded. Moreover, as we have discussed, the quality of data begins at the point of collection and can be further affected through data migration and data extraction. Educating users on where, how, and why data can be lost or altered may provide valuable insight and inform researchers about its limitations.

Given our findings regarding discrepancies in race/ethnicity classifications, particularly under the "Unknown" and "Other" categories, procedures related to ETL processes may be refined. For example, updates to EHR race/ethnicity classification should prompt updates to ETL rules and codes to map appropriately to external databases. This type of improvement may be incorporated into scheduled reviews of the ETL codes and categorization or as part of the organization's governance process. Another approach would be the implementation of dynamic ETL, a combined

automated and manual coding, a scalable solution that has been shown to improve harmonization.[31] Additionally, involving stakeholders, interested in the extracted data, into the mapping process may improve harmonization and increase data completion and correctness.[9]

Conclusion

While differences were found in the aggregate output between a self-service exploration tool and a CDW whose data are sourced from a single EHR, we believe that these variances do not prevent its utilization. As end users are increasingly provided the opportunity to extract data for various purposes, user training would ensure more accurate reporting. Additionally, as the data flow from its source to the end user, and as multiple touch points may affect data quality, user education on the underlying data model and subsequent data flow may provide an understanding of the etiology of such variances and instill a greater confidence in its reporting and use.

Clinical Relevance Statement

Discrepancies in aggregate patient data found between two sources with data originating from the same EHR reiterate the importance of user training and education on data extraction and data flow and prompt further discussions to improve data quality including possible improvements to ETL processes.

Multiple-Choice Questions

1. What is the primary function of an ETL process in relation to databases?
   a. To ensure database transactions are processed efficiently and quickly
   b. To analyze and visualize data for business intelligence purposes
   c. To migrate data from one database to another while performing necessary data transformations
   d. To secure databases against unauthorized access and data breaches

   Correct Answer: The correct answer is option c. To migrate data from one database to another while performing necessary data transformations. The primary function of an ETL process is to migrate data from one database to another while performing necessary data transformations. The ETL process involves three main steps:
   1. Extract: Retrieving data from various source systems.
   2. Transform: Cleaning, enriching, and transforming the data into a suitable format or structure for analysis.
   3. Load: Loading the transformed data into the target database or data warehouse.

   ETL is essential for integrating data from multiple sources, ensuring data quality, and preparing data for analysis and reporting.

2. Which of the following is one of the primary benefits of extracting data from a data warehouse rather than through a graphical user interface within the electronic medical record?
   a. Improved user experience with intuitive interfaces
   b. Enhanced data security and access control
   c. Data warehouse may include the integration of additional data sources
   d. Real-time updates and data entry capabilities

   Correct Answer: The correct answer is option c. Data warehouse may include the integration of additional data sources. One of the primary benefits of extracting data from a data warehouse is that it often includes the integration of additional data sources. Data warehouses are designed to consolidate data from various systems and sources, providing a comprehensive and unified view of information. This integration allows for more robust data analysis and reporting capabilities, as it combines data from different parts of an organization, such as clinical, financial, and operational.

3. Which of the following is the least likely reason for discrepancies between actual patient information and the data extracted from a database?
   a. Human data entry errors
   b. Data integration and synchronization issues
   c. System downtimes during data extraction
   d. Differences in data formatting and standards

   Correct Answer: The correct answer is option c. System downtimes during data extraction. Discrepancies between actual patient information and the data extracted from a database can arise due to various factors. Human data entry errors (a.) can lead to incorrect information being recorded. Data integration and synchronization issues (b.) can cause mismatches if different systems are not properly aligned. Differences in data formatting and standards (d.) can result in data being misinterpreted or improperly displayed. However, system downtimes during data extraction (c.) are the least likely reason for such discrepancies. While downtimes can affect the availability of data or delay its extraction, they do not inherently cause inaccuracies in the data itself. The primary concern during downtimes is accessibility rather than the accuracy of the data.

Protection of Human and Animal Subjects

The studies were performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects. This study did not constitute human subject research and met the criteria for NHSR self-determination at the University of California, Irvine, CA.

Funding

None.

Conflict of Interest

None declared.

Acknowledgments

The authors would like to acknowledge the invaluable insight and expertise of Eliza Ali, Kimberly Ruprecht, Kathy Pickell, Nora Lewin, Andrea Hwang, and Peyton Politewicz.

References

1 Chishtie J, Sapiro N, Wiebe N, et al. Use of Epic electronic health record system for health care research: scoping review. J Med Internet Res 2023;25:e51003
2 Saini V, Jaber T, Como JD, et al. 623. Exploring 'Slicer Dicer', an extraction tool in EPIC, for clinical and epidemiological analysis. Open Forum Infect Dis 2021;8:S414–S415
3 Baughman DJ, Jabbarpour Y, Westfall JM, et al. Comparison of quality performance measures for patients receiving in-person vs telemedicine primary care in a large integrated health system. JAMA Netw Open 2022;5(09):e2233267
4 Bui R, Kasabali A, Dewan K. A retrospective analysis of COVID-19 tracheostomies: early versus late tracheostomy. Laryngoscope Investig Otolaryngol 2023;8(05):1154–1158
5 Shermon S, Fazio KM, Shim R, Abd-Elsayed A, Kim CH. Prescription trends in complex regional pain syndrome: a retrospective case-control study. Brain Sci 2023;13(07):1012
6 van der Lei J. Use and abuse of computer-stored medical records. Methods Inf Med 1991;30(02):79–80
7 Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013;20(01):144–151
8 Köpcke F, Trinczek B, Majeed RW, et al. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. BMC Med Inform Decis Mak 2013;13:37
9 Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016;4(01):1244
10 AbuHalimeh A. Improving data quality in clinical research informatics tools. Front Big Data 2022;5:871897
11 Tung TH, DeLaurentis P, Yih Y. Uncovering discrepancies in IV vancomycin infusion records between pump logs and EHR documentation. Appl Clin Inform 2022;13(04):891–900
12 Lee SJ, Grobe JE, Tiro JA. Assessing race and ethnicity data quality across cancer registries and EMRs in two hospitals. J Am Med Inform Assoc 2016;23(03):627–634
13 Mohamed Y, Song X, McMahon TM, et al; Greater Plains Collaborative. Electronic health record data quality variability across a multistate clinical research network. J Clin Transl Sci 2023;7(01):e130
14 Edmondson ME, Reimer AP. Challenges frequently encountered in the secondary use of electronic medical record data for research. Comput Inform Nurs 2020;38(07):338–348
15 Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible sources of bias in primary care electronic health record data use and reuse. J Med Internet Res 2018;20(05):e185
16 Ancker JS, Shih S, Singh MP, Snyder A, Edwards A, Kaushal RHITEC investigators. Root causes underlying challenges to secondary use of data. AMIA Annu Symp Proc 2011;2011:57–62
17 Kornegay C, Segal JB. Chapter 8: Selection of Data Sources. In: Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM, eds. Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. Rockville (MD): Agency for Healthcare Research and Quality (US); 2013. Accessed October 25, 2024 at: https://www.ncbi.nlm.nih.gov/books/NBK126195/
18 Wiley KK, Mendonca E, Blackburn J, Menachemi N, Groot M, Vest JR. Quantifying electronic health record data quality in telehealth and office-based diabetes care. Appl Clin Inform 2022;13(05):1172–1180
19 Klinger EV, Carlini SV, Gonzalez I, et al. Accuracy of race, ethnicity, and language preference in an electronic health record. J Gen Intern Med 2015;30(06):719–723
20 Magaña López M, Bevans M, Wehrlen L, Yang L, Wallen GR. Discrepancies in race and ethnicity documentation: a potential barrier in identifying racial and ethnic disparities. J Racial Ethn Health Disparities 2016;4(05):812–818
21 Samalik JM, Goldberg CS, Modi ZJ, et al. Discrepancies in race and ethnicity in the electronic health record compared to self-report. J Racial Ethn Health Disparities 2023;10(06):2670–2675
22 Cook LA, Sachs J, Weiskopf NG. The quality of social determinants data in the electronic health record: a systematic review. J Am Med Inform Assoc 2021;29(01):187–196
23 Johnson JA, Moore B, Hwang EK, Hickner A, Yeo H. The accuracy of race & ethnicity data in US based healthcare databases: a systematic review. Am J Surg 2023;226(04):463–470
24 Cook L, Espinoza J, Weiskopf NG, et al; N3C Consortium. Issues with variability in electronic health record data about race and ethnicity: descriptive analysis of the National COVID Cohort Collaborative Data Enclave. JMIR Med Inform 2022;10(09):e39235
25 Wang K, Grossetta Nardini H, Post L, Edwards T, Nunez-Smith M, Brandt C. Information loss in harmonizing granular race and ethnicity data: descriptive study of standards. J Med Internet Res 2020;22(07):e14591
26 Voss EA, Makadia R, Matcho A, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc 2015;22(03):553–564
27 Yin AL, Guo WL, Sholle ET, et al; Weill Cornell COVID-19 Data Abstraction Consortium. Comparing automated vs. manual data collection for COVID-specific medications from electronic health records. Int J Med Inform 2022;157:104622
28 Torres FBG, Gomes DC, Hino AAF, Moro C, Cubas MR. Comparison of the results of manual and automated processes of cross-mapping between nursing terms: quantitative study. JMIR Nurs 2020;3(01):e18501
29 Mohamed Y, Song X, McMahon TM, et al. Tailoring rule-based data quality assessment to the Patient-Centered Outcomes Research Network (PCORnet) Common Data Model (CDM). AMIA Annu Symp Proc 2023;2022:775–784
30 Rungvivatjarus T, Chong AZ, Patel A, Khare M, Bialostozky M, Kuelbs CL. Training pediatric physicians and staff to obtain data from the electronic health record. Healthcare (Amst) 2024;12(01):100733
31 Ong TC, Kahn MG, Kwan BM, et al. Dynamic-ETL: a hybrid approach for health data extraction, transformation and loading. BMC Med Inform Decis Mak 2017;17(01):134