J Diabetes Clin Res. 2026;8(1):58–63.

Journal of Diabetes and Clinical Research Review Article

Data Retrieval from the Electronic Health Record Requires Validation: New Challenges in the Era of ICD-10 and Big Data

Lisa M. Scheid, MD, MS[1,*], Charles R. Rosenfeld, MD[1]

[1]Department of Pediatrics, Division of Neonatal-Perinatal Medicine, University of Texas Southwestern Medical Center, Dallas, TX, United States

[*]Correspondence should be addressed to Lisa M. Scheid, lisa.scheid@utsouthwestern.edu

Received date: August 28, 2026, **Accepted date:** November 11, 2026

Citation: Scheid LM, Rosenfeld CR. Data Retrieval from the Electronic Health Record Requires Validation: New Challenges in the Era of ICD-10 and Big Data. J Diabetes Clin Res. 2026;8(1):58–63.

Copyright: © 2026 Scheid LM, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Background: As the availability of Electronic Health Records (EHR) has become more ubiquitous, data derived from the EHR has increased in parallel. Data retrieval was originally limited by the need for manual extraction; today, many vendors offer ways to quickly extract large volumes of electronically recorded data, opening doors to the assessment of large study populations. In 2019, we identified where errors can exist within data electronically extracted from the chart and identified several approaches that investigators should address in order to validate a subset of extracted variables and to determine their fitness for research inclusion. A review of twenty recently published manuscripts of retrospective research utilizing electronically extracted health record data showed only two (10%) documented data validation procedures by the researchers in the methods section, and only one (5%) documented data quality measures used by the large, federated, de-identified database from which their data was retrieved. We present an updated review of current practices in the use of the EHR for research, highlighting emerging challenges and the importance of transparent reporting of data validation and quality assessment methods.

Conclusion: The expansion of ICD codes and creation of large, federated databases have added additional complexities to the use of EHR for research purposes. Validation of electronically extracted data from health records should be required and the details of the validation processes clearly documented in published methods. When large de-identified databases are utilized, the platform's intrinsic data-quality assessment processes should be identified and reported. Currently, few manuscripts address these practices.

Keywords: Hypoglycemia, Gestational diabetes, Diabetic complications, Data, Validation, Research, Extraction, Electronic health records

Introduction

Nine years ago, we [1] embarked on a study of neonatal hypoglycemia using electronically extracted data from the Electronic Health Record (EHR). Within the data collection, we included an embedded validation study that compared the results of blinded manual data extraction to the electronic data extraction. Results of the validation study identified varying degrees of agreement with 43% (10/23) of extracted categorical variables having a kappa (*K*) coefficient for inter-user agreement of < 0.60, indicating only moderate to slight agreement. Notably, the subject of the study, "hypoglycemia," had the lowest *K* coefficient, 0.16. Although all patients had laboratory-confirmed hypoglycemia, defined as a serum glucose < 40 mg/dL, only 29% were correctly identified by ICD-9 code. Of the eight categorical variables with the worst *K* coefficient, five showed significant improvement after modifying the definition of the ICD-9 extraction code, while three variables, including "hypoglycemia," had *K* coefficients that could not be rectified. At the time of publication, we speculated as to whether the use of the new ICD-10 codes would improve extraction accuracy, since we discovered some variables that shared ICD-9 codes with other diagnoses leading to over identification.

International Classification of Diseases (ICD) codes are not the only variables that require validation, as was proven by our previous publication. It was also discovered that serum glucose measurements were recorded in three different locations within the EHR and point-of-care glucose results were, at times, duplicated with the same date and time stamp, which resulted in initial poor validity of only 33% for

this variable. Notably, when these issues were identified and corrected for, the extraction code validity increased to 84%. Importantly, these issues would not have been identified without our decision to validate the dataset [1].

The integration of the EHR into most large healthcare organizations was motivated by the American Recovery and Reinvestment Act of 2009. Since then, researchers have utilized these data for secondary purposes. The recent adoption of ICD-10 and large, federated databases for research pose new challenges with data validation. Given what we know about potential errors in electronically recorded data, it is unclear if present day investigators have validated their data derived from the EHR before publishing research results and conclusions. This review explores new challenges with use of electronic health data for research and the importance of transparent reporting of data validation and quality assessment methods.

ICD-9 to ICD-10

The transition from ICD-9 to ICD-10 resulted in a dramatic increase in diagnostic granularity and the number of diagnostic codes. According to Hedegaard et al. there were ~14,000 ICD-9-CM codes and ~69,000 ICD-10-CM codes in 2017, a five-fold increase [2]. **Table 1** illustrates the increase in ICD-9 vs. ICD-10 codes for specific variables used in our original validation study [3,4]. While this can lead to improvement in over-identification caused by nonspecific coding definitions, there is also the potential to under-identify due to increased extraction coding complexity and thus, incomplete or inaccurate code lists. In 2022, Schaefer et al. compared ICD-10 codes for seven common comorbid conditions in the EHR for COVID-19 patients with results from a manual chart review [5]. They reported that the sensitivity for comorbidity detection via ICD-10 code to be highly variable, depending on the location within the EHR and the timing of data retrieval relative to the time of hospitalization. For instance, the ICD-10 code for chronic kidney disease via the past medical history on the day of admission only had a sensitivity of 0.3. The sensitivity of chronic kidney disease via the past medical history increased to 0.44 after removing the time constraint of the day of admission and allowing for any hospitalization day. Notably, the same diagnosis obtained from the active problem list on the day of admission had a sensitivity of 0.68, which increased to 0.77 with removal of the time constraint, further highlighting that the location and timing of data extraction within the chart can produce widely variable results, especially

Table 1. Comparison of International Classification of Diseases (ICD-9) and ICD-10 codes for diagnostic variables.

Diagnosis	ICD-9-CM Codes	Total ICD-9 Number	ICD-10-CM Codes	Total ICD-10 Number	Fold change
Hypoglycemia	251.0–251.2 775.6	4	E15 E16.0–16.2 P70.3–70.4	6	1.5x
Type 1 DM	250.#1, 250.#3*	20	E10†	87	4.4x
Type 2 DM	250.#0, 250.#2*	20	E11†	87	4.4x
Gestational DM	648.80–648.84	5	O24.410, O24.414–415, O24.419–420, O24.424–425, O24.429–430, O24.434–435, O24.439 O24.811–813, O24.819 O24.911–913 O24.919	20	4.0x
Infant of Diabetic Mother	775.0	1	P70.0–70.1	2	2.0x

DM: Diabetes Mellitus

*# denotes numbers 0–9, i.e. 250.#1 = 250.01, 250.11, 250.21,…, 250.91.

†See supplement for complete list as of August 2026.

for ICD codes. Consequently, the authors concluded that *all researchers need to validate extracted data* against a manual chart review. In addition, they called for standardization in data extraction methodologies and reporting practices for studies using ICD-10-based data [5].

Have Validation Practices Improved?

In 2009, Dean et al. reported that only 24% of published studies using EHR-derived data had documented validation of the EHR data with other sources of data [6]. At the time, EHR data was relatively new and the reliability questionable. By 2019, the availability of EHR-derived data was more common and, depending on the vendor, had some ability for electronic extraction. Based on our 2019 validation study [1], we also concluded that all researchers using electronically extracted data from the EHR needed to validate a subset of each variable in the dataset using manual chart review in order to identify potential sources of error and determine which variables are sufficiently accurate to remain in the dataset and those that should be removed [1]. The question today is whether validation practices have improved.

To investigate this, we examined the inclusion of validation practices by assessing 20 publications available on PubMed that utilized electronically extracted EHR data and were published in 2025 (**Figure 1**). Query language included "retrospective" and variations of "electronic health records" or "electronic medical records," but excluded "machine learning", "artificial intelligence" and "natural language processing" to focus results on clinical research and not informatics techniques. The dates of publication were from 1/1/2025–12/31/2025. A preliminary search yielded 3746 publications, illustrating the significant impact EHR data has had on the volume of clinical research in the last decade. Starting with the most recent publication, manuscripts were assessed for access to the full article and evidence of *electronic* data extraction in the methods section until we identified 20 inclusive manuscripts. Of note, 19 manuscripts were excluded due to: (1) inability to access the full article, (2) documentation of manual chart review instead of electronic data extraction, or (3) lack of manual vs. electronic data extraction distinction with a study population small enough to make manual extraction possible. Studies that lacked documentation distinguishing between manual vs. electronic extraction but had sufficiently large study populations to make manual data extraction infeasible (i.e., 50,000) were included. For included manuscripts, the methods section of each publication was analyzed for any documentation of data validation. Of the 20 manuscripts reviewed, only two (10%) included documented data validation. Ten (50%) were studies conducted at single centers or single healthcare organizations with access to source data but without any mention of data validation. The remaining eight (40%) utilized large, federated, de-identified databases which eliminates access to source data and thus makes source validation impossible.

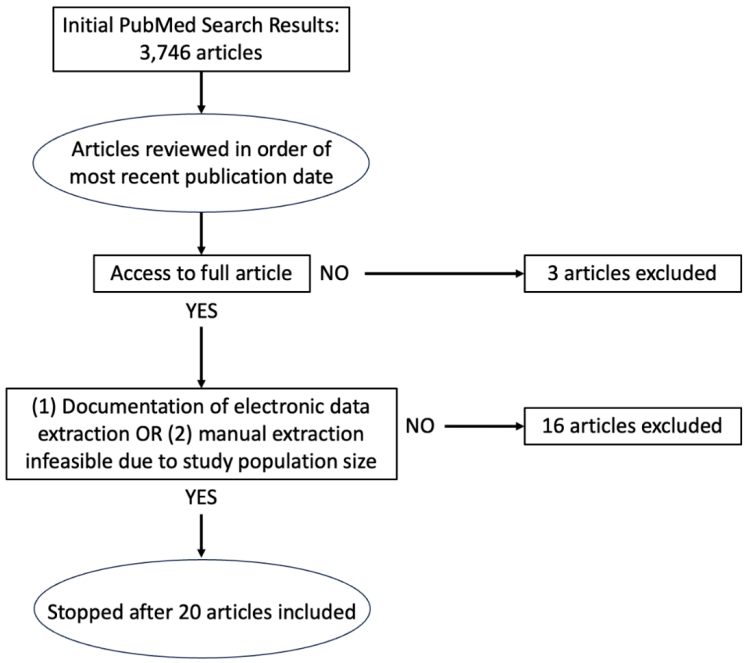


Figure 1. Flowchart of article selection for review of validation documentation.

Additionally, only one of the eight reported quality control measures used by the database (**Table 2**) [7–26]. In summary, this descriptive analysis highlights the infrequent inclusion of data validation practices in published research using electronic health data, raising uncertainty as to whether this crucial step was performed.

Big Data: When Source Records Are No Longer Available

Large, federated databases are increasingly available and offer unique research opportunities but also pose unique research challenges. For instance, TriNetX, the largest existing federated healthcare database, announced in June 2026 that it has surpassed 300 million individual patient files from over 240 healthcare organizations worldwide [27]. Such expansive data repositories can open doors for population-level research studies and are being utilized at an increasing rate. Notably, of the twenty manuscripts surveyed in **Table 2**, four of the eight (50%) that utilized large databases used data from TriNetX [19,20,25,26]. One challenge with these large databases is they often provide investigators only de-identified patient data, thus, making source data validation impossible. This, however, does not eliminate the obligation to evaluate and determine data validity. It merely changes how assessment of data quality should be performed. Kahn et al, published a Harmonized Data Quality Assessment Framework of completeness, conformance and plausibility. *Completeness* is the degree to which an expected data variable is available in the dataset and not missing. For example, a dataset with only 2% of the variable missing would be considered to have good completeness for that particular variable. *Conformance* is the degree to which the data adheres to a predefined format, data type and standard for each variable. For example, if a dataset has defined serum glucose units as mg/dL and records data as mmol/L this is considered poor conformance. Additionally, having some serum glucose values recorded as < 10 and others as actual integers (i.e. 3) is poor conformance due to lack of consistency. Lastly, *plausibility* is the degree to which the data is believable. For instance, a serum glucose value of 1500 mg/dL would be considered implausible [28]. This framework is the foundation for internal data quality standards claimed by both TriNetX and PCORnet (The National Patient-Centered Clinical Research Network), another large database with over 50 million patients [29,30] and should be utilized by researchers to assess EHR data quality for secondary use when access to source data is unavailable.

Understanding the internal data quality standards of large de-identified databases is essential and should be mandatory before trusting data accuracy; therefore, transparency regarding quality control measures is critical to establishing confidence in the data used. Demonstrating best practices, Cosmos, a large, federated database maintained by Epic Systems Corporation publishes completeness, conformance and plausibility metrics for variables as new data are incorporated in the database [31]. While no pre-defined acceptable metric threshold exists, this level of transparency allows researchers to determine for themselves whether each variable is fit for their own published research.

Expanding on data quality concepts, Feder described additional ways to assess data accuracy, including data triangulation with outside sources such as comparing statistical measures for research variables against the same statistical measures already published in the literature using identified patient data [32]. Large discrepancies call into question the validity of the de-identified data.

Table 2. Presence or lack of data validation documented in twenty manuscripts published electronically or in press in 2025.

Validation Discussed (n=2)	No Validation Discussed (n=10)	Direct source-chart validation not available (n=8)
Burger et al. [7]	Mirza et al. [9]	Kim et al. [19]
Zaboli et al. [8]	Kim et al. [10]	Vesole et al. [20]
	Omole et al. [11]	Shahnam et al. [21]
	Epstein et al. [12]	Bailey et al. [22]
	Cheeks et al. [13]	Einbinder et al. [23]
	Saey et al. [14]	Li et al. [24]
	Witte et al. [15]	Chang et al. [25]
	Jutras et al. [16]	Wang et al. [26]*
	Moore et al. [17]	
	Sorin et al. [18]	

*Reported database internal quality control measures in the methods section.

Recommendations for Researchers

Data derived from the EHR and used for the secondary purpose of research must be held to the same high standard as any other research database. Importantly, investigators are obligated to validate all data used for clinical conclusions as they may lead to unfounded changes in clinical practice. When source data are available, a subset of each variable electronically extracted from the EHR should be compared against a manual chart review to determine potential sources of error and which variables can reliably remain in the study for analyses. Data quality frameworks offer additional tools to assess data fitness, which become increasingly more important when source data are unavailable due to use of de-identified data from large databases. As noted above, understanding internal data quality control measures used in large databases is crucial to instilling trust of the data analyzed; therefore, it is important to only use data from sources with transparency in internal data control measures, such as conformance, completeness and plausibility, as outlined by Kahn et al [28]. Importantly, validation and data quality methods used by all investigators and in all databases should be documented within published works to instill trust in the reader [32,33].

Recommendations for Reviewers and Journals

Clinical journals have both the privilege and responsibility of disseminating new and innovative research and the conclusions derived therein. As a result, it is recommended that journals should require a description of data validation or quality assessment procedures within the methods section for all studies utilizing data derived from EHR or large healthcare databases to ensure their readers that substantial effort has been made to minimize data errors.

Recommendation for Readers

Scientists, students, investigators, and clinicians rely on published research to make informed decisions regarding whether a change in clinical practice is warranted or should be considered. It is, therefore, important for all readers to understand the limitations and errors that can exist within data electronically extracted from the EHR and to critically appraise the quality of the data used by looking for validation processes within the methods section of published research. In the case of data used from large, federated databases, both authors and readers should familiarize themselves with the data quality assessment measures published or unpublished by the database owners and determine for themselves whether sufficient evidence exists to support confidence in the reliability of the data.

Conclusion

Data electronically extracted from clinical health records still require validation from source data and the details documented in published methods. When large de-identified databases are utilized, the platform's intrinsic data quality processes should be reported along with additional efforts by researchers to validate the dataset.

Conflict of Interest

The authors declare that they have no conflicts of interests.

Funding Statement

The authors have no funding support to declare.

References

1. Scheid LM, Brown LS, Clark C, Rosenfeld CR. Data electronically extracted from the electronic health record require validation. J Perinatol. 2019 Mar;39(3):468–74.
2. Hedegaard HB, Johnson RL, Ballesteros MF. Proposed ICD-10-CM Surveillance Case Definitions for Injury Hospitalizations and Emergency Department Visits. Natl Health Stat Report. 2017 Jan;(100):1–8.
3. The web's free ICD-9-CM and ICD-10-CM medical coding reference [Internet]. Alkaline Software; 2006 [cited 2026 Aug 23]. Available from: https://www.icd9data.com
4. The web's free 2026 ICD-10-CM/PCS medical coding reference [Internet]. Alkaline Software; 2010 [cited 2026 Aug 23]. Available from: https://www.icd10data.com
5. Schaefer JW, Riley JM, Li M, Cheney-Peters DR, Venkataraman CM, Li CJ, et al. Comparing reliability of ICD-10-based COVID-19 comorbidity data to manual chart review, a retrospective cross-sectional study. J Med Virol. 2022 Apr;94(4):1550–7.
6. Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. Review: use of electronic medical records for health outcomes research: a literature review. Med Care Res Rev. 2009 Dec;66(6):611–38.
7. Burger J, Hess R. Impact of Pharmacist Consultations in Management of Type 1 Diabetes in a Primary Care Clinic: A Retrospective Cohort Study. J Pharm Pract. 2026 Oct;39(5):387–91.
8. Zaboli A, Ziller M, Brigo F, Cipriano A, Pfeifer N, Voza A, et al. The impact of chronic conditions on emergency department length of stay. Intern Emerg Med. 2026 Aug;21(5):1797–805.
9. Mirza M, Kim H, Pilat C. Does Language Proficiency Affect Completion of Telerehabilitation Video Visits? A Retrospective Study. J Immigr Minor Health. 2026 Aug;28(4):863–9.
10. Kim Y, Kang J, Song J, Park J, Bae S, Kim B, et al. Evaluation of the Real-World Practice of Model-Informed Therapeutic Drug Monitoring and Its Association With Pharmacokinetic Target Attainment Rates: A Large-Scale, Single-Center Retrospective Study. Ther Drug Monit. 2026 Aug 1;48(4):451–62.
11. Omole O, Palin V, Watson K, Myers J. Examining the Magnitude

of Maternal Ethnic and Socioeconomic Inequalities on Foetal Growth Restriction and Preterm Birth: A Cohort Study Set in North West England. J Racial Ethn Health Disparities. 2026 Aug;13(4):2535–44.

12. Epstein RH, Dexter F, Gayer S. The Influence of Low-Dose Dexmedetomidine on Recovery Times Following Sedation for Vitreoretinal Surgery: A Retrospective Cohort Study. Anesth Analg. 2026 Jul 1;143(1):174–81.

13. Cheeks M, Lee D, Su S, Andebrhan S, Lancki N, Bretschneider CE. Baseline Anticholinergic Medication Burden in Patients with Overactive Bladder. Int Urogynecol J. 2026 Jul;37(7):2027–33.

14. Saey SA, Halvorsen AJ, Nordhues HC. Gender Differences in Inpatient Secure Chat Messaging Burden Among Internal Medicine Residents. J Gen Intern Med. 2026 Jul;41(9):2388–94.

15. Witte M, McDermott C, Levy KL, Nelsen K, Shrestha S, Tian S, et al. Gender differences in secure chat and electronic health records use among hospital-based physicians. J Hosp Med. 2026 Jul;21(7):715–23.

16. Jutras G, Fenton C, Flemming JA, Gamez J, Irani RA, Kushner T, et al. Epidemiologic Trends in Acute, Chronic, and Pregnancy-Specific Liver Diseases in Pregnancy. Am J Gastroenterol. 2026 Jul 1;121(7):1630–37.

17. Moore KS, Colbert A, Zoucha R, Hendricks-Ferguson V, Taylor J. Race, Ethnicity, Neighborhood Disparities and Palliative Care Utilization: Retrospective Cohort Study. Am J Hosp Palliat Care. 2026 Jul;43(7):736–44.

18. Sorin V, Korfiatis P, Langer SG, Hahn LD, Bratt AK, Cook CJ, et al. Factors Impacting the Performance of Deep Learning Detection of Pulmonary Emboli. J Am Coll Radiol. 2026 Jun;23(6):965–72.

19. Kim SH, Marshall RF, Wai KM, Rahimy E, Vingopoulos F, Mishra K, et al. Ocular Adverse Events of Antibody Drug Conjugates and Association With Survival From a Real-World Database. Cornea. 2026 Sep 1;45(9):1145–50.

20. Vesole AS, Gordon SA, Sedaghat AR, Phillips KM, Forbes J, Andaluz N, et al. Reduction of Cerebrospinal Fluid Leaks in Idiopathic Intracranial Hypertension with Glucagon-like Peptide-1 Receptor Agonist Use. J Neurol Surg B Skull Base. 2025 Aug 14;87(4):490–94.

21. Shahnam A, Lewin J, Hamilton A, Luen SJ, Desai J, Orme L, et al. Real-World Trends and Outcomes for Localized High-Grade Soft Tissue Sarcoma: Insights from a Sarcoma Unit in Australia. Asia Pac J Clin Oncol. 2026 Aug;22(4):608–17.

22. Bailey SR, Hwang J, Lucas JA, Lyon-Scott K, Marino M, Datta R, et al. Tobacco Use Assessment and Cessation Medication Orders Among Primary Care Patients by Ethnicity, Sex, and Acculturation Indicators. J Gen Intern Med. 2026 Aug;41(11):3121–29.

23. Einbinder Y, Hornik-Lurie T, Cohen-Hagai K, Benchitrit S, Kalantar-Zadeh K, Karpati T. Comparative Patterns in CKD among Native Israeli and Immigrant Ethnic Populations. Clin J Am Soc Nephrol. 2026 Jul 1;21(7):1269–77.

24. Li X, Suh K, He N, Karimzadeh I, Kane-Gill SL. Remdesivir and Renal Outcomes in Hospitalized COVID-19 Patients Without Severe Renal Impairment: A Retrospective Cohort Study. Ann Pharmacother. 2026 Jul;60(7):667–76.

25. Chang HC, Jhang YS, Chen SJ, Gau SY. Alopecia Areata and Risk of Antiphospholipid Syndrome: A Multicenter Retrospective Cohort Study Using Global Electronic Medical Records. J Cutan Med Surg. 2026 Jul-Aug;30(4):431–33.

26. Wang PR, Mascha EJ, Li Y, Bobak L, Dorney I, Costa G, et al. Impact of X-Waiver Elimination on Post-Hospital Discharge Buprenorphine Prescribing: An Interrupted Time Series Analysis. J Addict Med. 2026 Jul-Aug 01;20(4):453–58.

27. Tunks K. TriNetX Live™ network surpasses 300 million patients, expanding decision-grade evidence across global healthcare [Internet]. 2026 [cited 2026 Aug 23]. Available from: https://trinetx.com/press-releases/trinetx-live-network-surpasses-300-million-patients-expanding-decision-grade-evidence-across-global-healthcare

28. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). 2016 Sep 11;4(1):1244.

29. Stein E, Hüser M, Amirian ES, Palchuk MB, Brown JS. TriNetX Dataworks-USA: Overview of a Multi-Purpose, De-Identified, Federated Electronic Health Record Real-World Data and Analytics Network and Comparison to the US Census. Pharmacoepidemiol Drug Saf. 2025 Sep;34(9):e70198.

30. PCORnet. Data scale, quality and security [Internet]. 2026 [cited 2026 Aug 23]. Available from: https://pcornet.org/data

31. Howat M, Martell S, Tarabichi Y, Kaelber DC, Chen SS, Freedman H. Advancing Real-World Evidence Generation: Growth and Lessons Learned from the Early Years of Cosmos. ACI open. 2026 Apr 24;10(1):e10–23.

32. Feder SL. Data Quality in Electronic Health Records Research: Quality Domains and Assessment Methods. West J Nurs Res. 2018 May;40(5):753–66.

33. Wang W, Ferrari D, Haddon-Hill G, Curcin V. Electronic health records as source of research data. In: Colliot O, editor. Machine learning for brain disorders [Internet]. New York (NY): Humana; 2023. p. 331–54.