Healthcare 8 (2020) 100444

Contents lists available at ScienceDirect

Healthcare

journal homepage: www.elsevier.com/locate/healthcare

How much can we trust electronic health record data?

Samuel T. Savitz[a], Lucy A. Savitz[b,*], Neil S. Fleming[c], Nilay D. Shah[d], Alan S. Go[a,e,f]

[a] Kaiser Permanente Northern California Division of Research, USA
[b] Kaiser Permanente Center for Health Research, USA
[c] Hankamer School of Business, Baylor University, USA
[d] Division of Health Care Policy & Research, The Mayo Clinic, USA
[e] Department of Epidemiology, Biostatistics and Medicine, University of California, San Francisco, USA
[f] Departments of Medicine, Health Research and Policy, Stanford University School of Medicine, USA

A B S T R A C T

Trust in EHR data is becoming increasingly important as a greater share of clinical and health services research use EHR data. We discuss reasons for distrust and acknowledge limitations. Researchers continue to use EHR data because of strengths including greater clinical detail than sources like administrative billing claims. Further, many limitations are addressable with existing methods including data quality checks and common data frameworks. We discuss how to build greater trust in the use of EHR data for research, including additional transparency and research priority areas that will both enhance existing strengths of the EHR and mitigate its limitations.

1. Background and motivation

With rapidly expanding use of electronic health record (EHR) data for clinical and health services research, some have questioned the quality of evidence generated by learning health systems. At a recent meeting of the National Academy of Medicine, selected editors from leading medical journals expressed concerns about the reliability and generalizability of research using EHR data.[1] The distrust may extend to other stakeholders including administrators, clinicians, and policymakers, and this distrust may be a barrier limiting the impact of EHR-based research on clinical practice. This skepticism raises important questions about why such doubt exists, whether the distrust is justified, and what can be done to improve trust.

2. Sources of distrust

Several characteristics about EHR-generated data may contribute to concerns about data quality. First, EHR data are often not optimized for research given they are collected primarily for clinical and administrative purposes.[2] However, this concern applies even more so to administrative claims, which are transactional, primarily collected for reimbursement and lack important clinical details.[3] Therefore, quality assurance is an important consideration with either EHR data that are optimized for documenting the clinical experience or administrative claims that are used primarily for billing.

Second, there are concerns about the accuracy of EHR documentation. Given the documentation burden required by clinicians and limited time available to complete charting,[4–6] occasionally, copying and pasting text can occur across notes for a patient even if the text no longer applies clinically.[7,8] Further, the same or conflicting information can be entered in multiple places that can vary between structured coded fields vs. unstructured notes.[9] Such findings can contribute to a perception of lower data quality.[10] While inaccurate documentation creates challenges for using EHR data, this limitation applies similarly or more so to administrative claims. Additionally, despite documentation issues, free-text fields have been a valuable source of clinical information not captured in structured codes.[11–13]

Third, greater confidence in billing claims over EHR data may reflect greater availability, familiarity, and belief that administrative codes are more standardized. Administrative claims have been pervasively used for research since the 1970's,[3] but there has been rapid growth and implementation of EHR systems following the 2009 HITECH Act.[14,15] Administrative claims, including frequently studied Medicare claims, are based on encounters linked to diagnostic and procedure codes associated with standard definitions. There is also a fairly standard data structure such that administrative claims from various payers look similar.[16] The familiarity and perceived standardization may lead to greater trust and willingness to accept key limitations including “code creep” and lack of accuracy across different practice settings.[17–20] In contrast to administrative claims, there is significant variation in EHR systems even though increasingly practices use only one of several major commercial vendors rather than homegrown systems. Nevertheless, completely standardized data models that promote further confidence in EHR data from multiple platforms are

* Corresponding author.
E-mail address: Lucy.A.Savitz@kpchr.org (L.A. Savitz).

https://doi.org/10.1016/j.hjdsi.2020.100444
Received 11 October 2019; Received in revised form 25 May 2020; Accepted 11 June 2020
Available online 08 July 2020

needed. Even across health systems using the same vendor, there can be varying data entry rules (e.g., acceptable values) and validation procedures (e.g., how unrealistic values are identified and addressed).[21,22] The lower degree of familiarity with and standardization of EHR data across different systems make it more challenging to ensure the quality of EHR data for research.

3. Limitations of EHR data

We reviewed published reports related to EHR data quality. The following potential limitations of EHR data emerged.[3]

1. Missing data: Incomplete data for patients receiving care from multiple healthcare systems.[21,23]
2. Errors: Incorrect data entered by clinicians [21,23] can contribute to bias or lack of precision.
3. Longitudinal inconsistencies: Inconsistency in data elements across time due to changes in data collection procedures, clinical workflow, EHR infrastructure, documentation practices, and data derived from ICD-based codes that have transitioned over time.[21–25]
4. Site and provider inconsistencies: Inconsistency caused by differences in how data are entered and organized.[21–24]
5. Unstructured data: While also considered a strength because of their contextual richness, data are unstructured in progress notes or reports potentially appearing in multiple places of the EHR, which adds to the challenge of data standardization.[21]
6. Ability to capture all clinical data: All relevant data may not be readily accessible (e.g., radiologic images, digital waveforms, and device specifications).

4. Strengths of EHR data

Despite potential limitations, EHR data offer notable advantages for research because of the rich clinical information that is unavailable in claims, such as vital signs, laboratory and pathology results, progress notes, procedure reports, and increasingly patient-reported measures.[26] While some studies use only EHR data or only administrative claims, other studies combine both data sources for the same patients to leverage potential strengths of each source.[27–29]

EHR data are generally more representative of the broader population than administrative claims that are limited to a single payor. Most EHR sources include patients from a range of payors including uninsured patients who receive care at a health system, and patients with multiple payors.[30] EHR data also typically have a much shorter lag time for access than administrative claims. For example, Medicare administrative claims have a lag of approximately two years,[31,32] while EHR data can be available in real-time or within a few days of collection.[32] EHR data can uniquely support actionable quality improvement and research studies given the role of the EHR in patient care and facilitating process measurement and clinical and patient-reported outcome data.[33,34]

5. What is being done to address limitations?

Researchers have developed tools and approaches to improve the quality of EHR data and transparency in reporting quality issues. First, this includes regularly performing data checks (audits) to document and address limitations, which may include assessing the amount of missing data, implausible values, expected relationships among variables, trends over time, and differences across systems or facilities.[21,22,35] This is frequently done in collaboration with clinical and IT professionals,[21,22,35] which is critical for understanding the data workflow and context in which information was entered into the EHR.[36] In cases such as missing diagnoses or medication use, the information may not be available from the EHR in any form and this issue is a limitation. However, many of the observed problems are addressable using alternative measures, triangulation of multiple measures for the same condition, standardizing inconsistencies like the use of different units, and using natural language processing (NLP) algorithms to derive structured, useable information from unstructured text.[21,22,35]

Second, distributed data networks standardize many EHR data elements across partnering health systems. For example, the Health Care Systems Research Network (HCSRN)[37] requires partner institutions to perform standardized data checks, mapping, definitions and formatting to a common data model to facilitate multi-institutional research. Examples of EHR data elements where this has been done include vital signs, body mass index and laboratory results, among others. Regular evaluation of HCSRN data across sites and over time within sites is performed to identify potential inconsistencies for sites to investigate and address.[37] Another example is the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM). OMOP CDM enables different types of observational data sources (EHR and administrative claims) to be combined into a single common data format using the same vocabulary, facilitating the use of the data for research and quality improvement efforts.[38]

Third, several tools have been developed to assist in reporting EHR data quality issues. Such tools enable readers to better understand data quality issues and the potential impact on results. The tools include quality assessments specific to EHR data[39–41] and an extension of the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines for routinely collected data including EHR data.[42] While these tools represent advances in how to report on data quality issues, no consensus exists on the effectiveness of any particular tool to date. Further, a review of studies using EHR data found that most studies used ad hoc approaches to assess data quality or did not report how data quality was assessed.[43]

6. What more can be done to build trust?

First, researchers should systematically report on EHR data quality and steps taken to optimize it. These efforts could incorporate quality assessment tools[39–41] and reporting guidelines.[42] Greater data quality transparency would assure readers that researchers have carefully considered potential limitations and made efforts to mitigate them.

Second, more research is needed to advance EHR data analysis. The Electronic Data Methods (EDM) Forum with support from the Agency for Healthcare Research and Quality brought together a diverse group of stakeholders to identify gaps and promote best practices for EHR research, including knowledge sharing through *eGEMs* (*Generating Evidence & Methods to Improve Patient Outcomes*) in 2013. In 2019, *eGEMs* transitioned to a section within *Healthcare: The Journal of Delivery Science and Innovation* to maintain a home for work that advances the practice of EHR research.[44] The EDM Forum and *eGEMs* identified areas where more research would improve EHR data quality including: 1) advances in the use and implementation of NLP to accurately classify and interpret unstructured data including addressing methodological challenges for NLP and quality issues for unstructured data, 2) HIPAA-compliant tracking of patient care received outside a health system to minimize missing data, 3) comparing the validity of different statistical approaches to missing data, 4) incorporating patient-reported outcomes to complement clinically-defined outcomes, and 5) enabling patients to view their data through OpenNotes [45–47] and report documentation errors.[46]

Third, researchers should work more closely with stakeholders including clinicians, administrators, and IT professionals to improve the quality of EHR data.[21,22,35] More work can be done with clinicians to support accurate documentation without eliminating flexibility and increasing documentation burden. For example, feedback reporting involves raising documentation issues with clinicians but does not necessarily make documentation more structured or burdensome.[21,48,49] Researchers should also remind clinicians that they are beneficiaries of

data quality improvement processes and research provides data quality 'stress testing'. By demonstrating the value of data quality to clinicians, they will be more likely to participate in these efforts. Further, researchers can partner with administrators and IT professionals to reduce documentation burden [4–6,50] and modify data fields to encourage documentation accuracy.

Finally, the ability to capture both complete and longitudinal data remains a challenge. This can be potentially achieved through HIPAA-compliant tools that enable patients to link their EHR data in a single place and share with researchers (e.g. Hugo PHR [51]). Tools such as these may become more feasible given the Office of the National Coordinator's new rules supporting interoperability and data exchange,[52] although additional work will be necessary to improve collaboration across healthcare systems and adequately address privacy and business confidentiality issues. However, these tools could facilitate greater completeness and longitudinal follow-up across multiple EMRs.

7. Conclusion

We have noted challenges in using EHR data for research that may lead to distrust and limit the impact of findings for policy or practice. We have, however, identified important strengths and demonstrated steps taken to address these challenges. As research use of EHR data by learning health systems expands, it is critical that we put this promising data source into context with respect to its considerable strengths in contrast to its addressable limitations.

Declaration of competing interest

The authors do not have any conflicts of interest to disclose. Dr. Savitz received funding from The Permanente Medical Group Delivery Science Fellowship Program that partially supported this work.

References

1. *NAM Leadership Consortium-Digital Health Learning and Clinical Effectiveness Research Collaboratives Meeting*. 2019; 2019https://nam.edu/event/nam-leadership-consortium-digital-health-learning-and-clinical-effectiveness-research-collaboratives-meeting/, Accessed date: 30 July 2019.
2. Nordo AH, Levaux HP, Becnel LB, et al. Use of EHRs data for clinical research: historical progress and current applications. *Learn Health Syst.* 2019;3:e10076.
3. Kari F, Bryan B, Paul J. The use of claims data in healthcare research. *Open Publ Health J.* 2009;2.
4. Arndt BG, Beasley JW, Watkinson MD, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. *Ann Fam Med.* 2017;15:419–426.
5. Sinsky C, Colligan L, Li L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Ann Intern Med.* 2016;165:753–760.
6. Tai-Seale M, Olson CW, Li J, et al. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. *Health Aff.* 2017;36:655–662.
7. Tsou AY, Lehmann CU, Michel J, Solomon R, Possanza L, Gandhi T. Safe practices for copy and paste in the EHR. Systematic review, recommendations, and novel model for health IT collaboration. *Appl Clin Inf.* 2017;8:12–34. https://doi.org/10.4338/aci-2016-09-r-0150.
8. Wang MD, Khanna R, Najafi N. Characterizing the source of text in electronic health record progress notes. *JAMA Internal Med.* 2017;177:1212–1213.
9. Palchuk MB, Fang EA, Cygielnik JM, et al. An unintended consequence of electronic prescriptions: prevalence and impact of internal discrepancies. *J Am Med Inf Assoc.* 2010;17:472–476.
10. Hoffman S, Podgurski A. Big bad data: law, public health, and biomedical databases. *J Law Med Ethics : J Am Soc of Law, Med; Ethics.* 2013;41(Suppl 1):56–60. https://doi.org/10.1111/jlme.12040.
11. Nelson SD, Lu CC, Teng CC, et al. The use of natural language processing of infusion notes to identify outpatient infusions. *Pharmacoepidemiol Drug Saf.* 2015;24:86–92. https://doi.org/10.1002/pds.3720.
12. Carrell DS, Cronkite D, Palmer RE, et al. Using natural language processing to identify problem usage of prescription opioids. *Int J Med Inf.* 2015;84:1057–1064. https://doi.org/10.1016/j.ijmedinf.2015.09.002.
13. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J Am Med Inf Assoc : JAMIA.* 2019;26:364–379. https://doi.org/10.1093/jamia/ocy173.
14. Blumenthal D. Launching hitech. *N Engl J Med.* 2010;362:382–385.
15. DesRoches CM, Charles D, Furukawa MF, et al. Adoption of electronic health records grows rapidly, but fewer than half of US hospitals had at least a basic system in 2012. *Health Aff.* 2013;32:1478–1485.
16. Mitchell JB, Bubolz T, Paul JE, et al. Using Medicare claims for outcomes research. *Med care.* 1994:JS38–JS51.
17. Steinwald B, Dummit LA. Hospital case-mix change: sicker patients or DRG creep? *Health Aff.* 1989;8:35–47. https://doi.org/10.1377/hlthaff.8.2.35.
18. Hsia DC, Krushat WM, Fagan AB, Tebbutt JA, Kusserow RP. Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. *N Engl J Med.* 1988;318:352–355. https://doi.org/10.1056/nejm198802113180604.
19. Silverman E, Skinner J. Medicare upcoding and hospital ownership. *J Health Econ.* 2004;23:369–389. https://doi.org/10.1016/j.jhealeco.2003.09.007.
20. Psaty BM, Boineau R, Kuller LH, Luepker RV. The potential costs of upcoding for heart failure in the United States. *Am J Cardiol.* 1999;84:108–109. https://doi.org/10.1016/s0002-9149(99)00205-2 a9.
21. Bayley KB, Belnap T, Savitz L, Masica AL, Shah N, Fleming NS. Challenges in using electronic health record data for CER: experience of 4 learning organizations and solutions applied. *Med care.* 2013;51:S80–S86.
22. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. *Med care.* 2013;51:S22–S29. https://doi.org/10.1097/MLR.0b013e31829b1e2c.
23. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med care.* 2013;51:S30–S37. https://doi.org/10.1097/MLR.0b013e31829b1dbd.
24. Estiri H, Stephens K, e)-v DQ(. *A Database-Agnostic Framework for Exploring Variability in Electronic Health Record Data Across Time and Site Location.* vol. 5. Washington, DC: EGEMS; 2017:3. https://doi.org/10.13063/2327-9214.1277.
25. Panozzo CA, Woodworth TS, Welch EC, et al. Early impact of the ICD-10-CM transition on selected health outcomes in 13 electronic health care databases in the United States. *Pharmacoepidemiol Drug Saf.* 2018;27:839–847. https://doi.org/10.1002/pds.4563.
26. D'Avolio LW, Farwell WR, Fiore LD. Comparative effectiveness research and medical informatics. *Am J Med.* 2010;123:e32–e37. https://doi.org/10.1016/j.amjmed.2010.10.006.
27. Kharrazi H, Chi W, Chang HY, et al. Comparing population-based risk-stratification model performance using demographic, diagnosis and medication data extracted from outpatient electronic health records versus administrative claims. *Med care.* 2017;55:789–796. https://doi.org/10.1097/mlr.0000000000000754.
28. Devoe JE, Gold R, McIntire P, Puro J, Chauvie S, Gallia CA. Electronic health records vs Medicaid claims: completeness of diabetes preventive care data in community health centers. *Ann Fam Med.* 2011;9:351–358. https://doi.org/10.1370/afm.1279.
29. Angier H, Gold R, Gallia C, et al. Variation in outcomes of quality measurement by data source. *Pediatrics.* 2014;133:e1676–e1682. https://doi.org/10.1542/peds.2013-4277.
30. Heintzman J, Marino M, Hoopes M, et al. Using electronic health record data to evaluate preventive service utilization among uninsured safety net patients. *Prev Med.* 2014;67:306–310.
31. Mues KE, Liede A, Liu J, et al. Use of the Medicare database in epidemiologic and health services research: a valuable source of real-world evidence on the older and disabled populations in the US. *Clin Epidemiol.* 2017;9:267–277. https://doi.org/10.2147/clep.S105613.
32. Doshi JA, Hendrick FB, Graff JS, Stuart BC. Data, data everywhere, but access remains a big issue for researchers: a review of access policies for publicly-funded patient-level health care data in the United States. *EGEMS (Washington, DC).* 2016;4:1204. https://doi.org/10.13063/2327-9214.1204.
33. Amarasingham R, Moore BJ, Tabak YP, et al. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Med care.* 2010;48:981–988. https://doi.org/10.1097/MLR.0b013e3181ef60d9.
34. Amarasingham R, Patel PC, Toto K, et al. Allocating scarce resources in real-time to reduce heart failure readmissions: a prospective, controlled study. *BMJ Qual Saf.* 2013;22:998–1005. https://doi.org/10.1136/bmjqs-2013-001901.
35. Welch G, von Recklinghausen F, Taenzer A, Savitz L, Weiss L. *Data Cleaning in the Evaluation of a Multi-Site Intervention Project.* vol. 5. Washington, DC: EGEMS; 2017:4. https://doi.org/10.5334/egems.196.
36. Johnson KE, Kamineni A, Fuller S, Olmstead D, Wernli KJ. How the provenance of electronic health record data matters for research: a case example using system mapping. *EGEMS (Washington, DC).* 2014;2:1058. https://doi.org/10.13063/2327-9214.1058.
37. Ross TR, Ng D, Brown JS, et al. The HMO research network Virtual data warehouse: a public data model to support collaboration. *EGEMS (Wash DC).* 2014;2:1049. https://doi.org/10.13063/2327-9214.1049.
38. *OMOP Common Data Model*. 2019; 2019https://www.ohdsi.org/data-standardization/the-common-data-model/, Accessed date: 6 October 2019.
39. Weiskopf NG, Bakken S, Hripcsak G, Weng C. *A Data Quality Assessment Guideline for Electronic Health Record Data Reuse.* . Washington, DC: EGEMS; 2017 5:14. 10.5334/egems.218.
40. 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 (Washington, DC).* 2016;4:1244. https://doi.org/10.13063/2327-9214.1244.
41. Sengupta S, Bachman D, Laws R, et al. *Data Quality Assessment and Multi-Organizational Reporting: Tools to Enhance Network Knowledge.* vol. 7. Washington, DC: EGEMS; 2019:8. https://doi.org/10.5334/egems.280.
42. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med.* 2015;12:e1001885https://doi.org/10.1371/journal.pmed.1001885.
43. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data

quality assessment: enabling reuse for clinical research. *J Am Med Inf Assoc : JAMIA*. 2013;20:144–151. https://doi.org/10.1136/amiajnl-2011-000681.
44. Wallace P. Moving ahead: what's next for the eGEMs community. *eGEMs*. 2019;7.
45. Walker J, Leveille S, Bell S, et al. OpenNotes after 7 Years: patient experiences with ongoing access to their clinicians' outpatient visit notes. *J Med Internet Res*. 2019;21:e13876https://doi.org/10.2196/13876.
46. Bell SK, Mejilla R, Anselmo M, et al. When doctors share visit notes with patients: a study of patient and doctor perceptions of documentation errors, safety opportunities and the patient-doctor relationship. *BMJ Qual Saf*. 2017;26:262–270. https://doi.org/10.1136/bmjqs-2015-004697.
47. Nazi KM, Turvey CL, Klein DM, Hogan TP, Woods SS. VA OpenNotes: exploring the experiences of early patient adopters with access to clinical notes. *J Am Med Inf Assoc : JAMIA*. 2015;22:380–389. https://doi.org/10.1136/amiajnl-2014-003144.
48. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *J Am Med Inf Assoc : JAMIA*. 2011;18:181–186. https://doi.org/10.1136/jamia.2010.007237.
49. Middleton B, Bloomrosen M, Dente MA, et al. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. *J Am Med Inf Assoc*. 2013;20:e2–e8.
50. Payne TH, Corley S, Cullen TA, et al. Report of the AMIA EHR-2020 Task Force on the status and future direction of EHRs. *J Am Med Inf Assoc*. 2015;22:1102–1110.
51. The future of health research has arrived. https://hugo.health/; 2019, Accessed date: 8 October 2019.
52. 21st century cures Act: interoperability, information blocking, and the ONC health IT certification program. *Office of the National Coordinator for Health Information Technology (ONC) DoHaHSH*. 2020; 2020.