Systematic Data Fitness Assessment Improves Validity and Replicability of Research Using Real-World Data

Hanieh Razzaghi, PhD, MPH [a,b], Kaleigh Wieand, BSc [b], Annabel Pinkney, MSLIS [b], Charles Bailey, MD, PhD [b,c]

[a] Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[b] Applied Clinical Research Center, Departments of Pediatrics and Biomedical and Health Informatics, Children's Hospital of Philadelphia, PA, USA
[c] Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

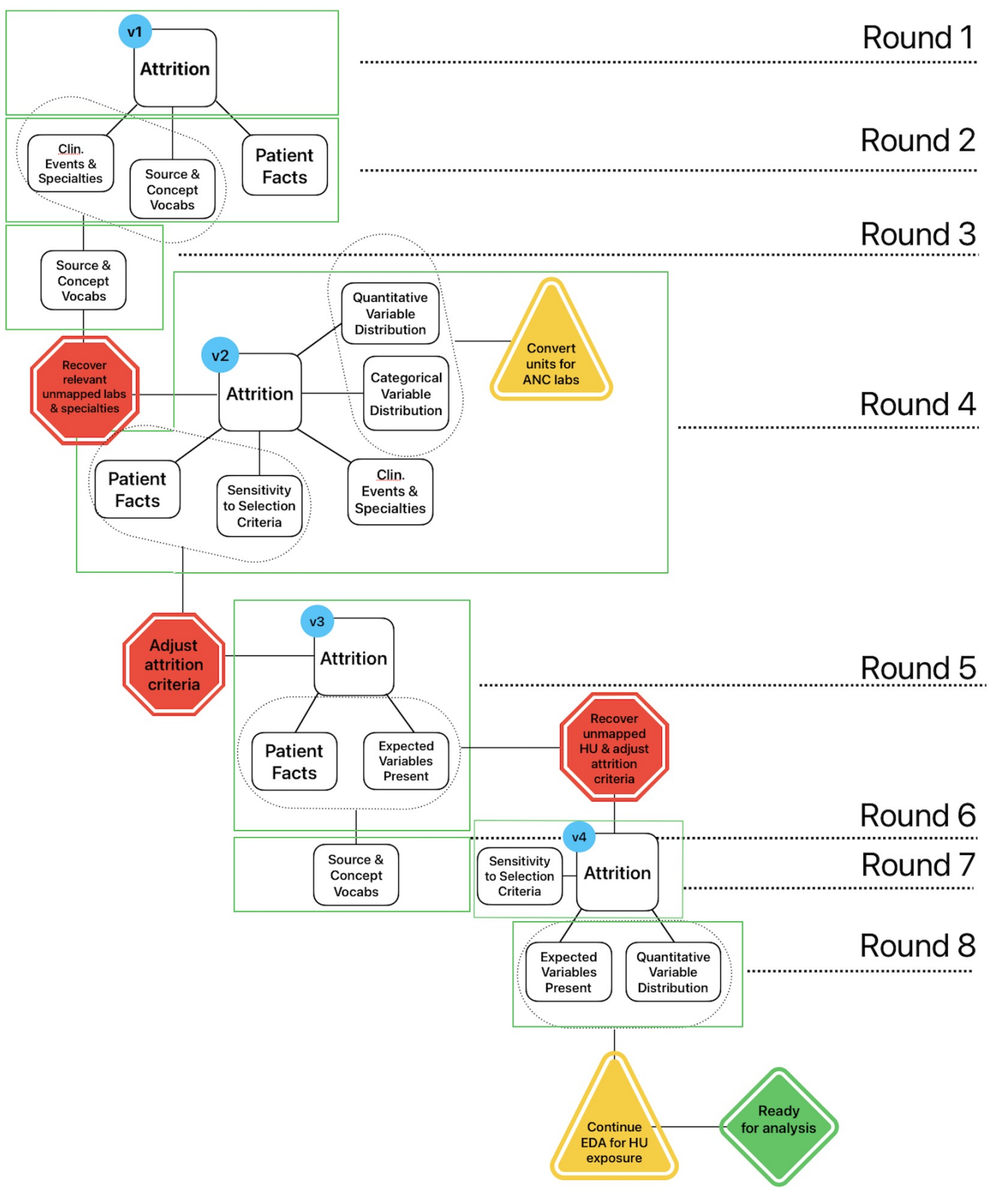
Background

Secondary use of electronic health record (EHR) data gives investigators the ability to conduct research and derive insights at a large scale for a lesser cost. These data are complex, so multi-institutional research networks often implement broad data quality (DQ) programs to ensure the data is generally usable and meet network governance standards[1,2]. However, study-specific needs are typically out of scope for these programs, leaving researchers with the responsibility of conducting ad hoc analyses to ensure the data is fit-for-use in the context of their study. Our team has developed a reusable, modularized tool to assess study-specific data quality (SSDQ), which is intended to reduce time spent developing ad hoc analyses and improve transparency and replicability in research[3]. Using this tool as a foundation, we demonstrate the utility of undertaking rigorous DQ assessment by replicating a study examining the impact of hydroxyurea treatment in a cohort of pediatric patients with the SS/Sβ^0 sickle-cell genotype[4,5].

Objective: Demonstrate the impact of standardized SSDQ assessment on research replicability using real-world EHR data generated in a multi-institutional pediatric research network

Methods

- The original study targeted patients < 18 years of age with the SS/Sβ^0 sickle-cell genotype that had been diagnosed and verified by hemoglobin typing. They evaluated the impact of hydroxyurea use on the following outcomes using a difference-in-difference approach:
  - Emergency department visits per age-year
  - Hospitalization days per age-year
  - Average annual hemoglobin per age-year
- We replicated the methods of the original study with two arms. The *AS-IS* arm was executed without any DQ intervention, while the *DFE* arm included 8 rounds of SSDQ evaluations using SQUBA to refine our cohort definition and capture of study variables.



Results

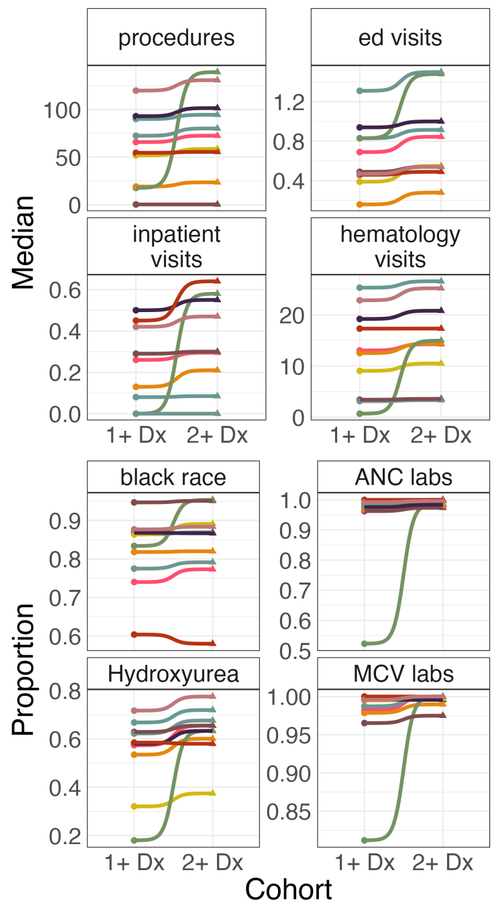
Cohort Descriptives

	CHOA N = 2,147	DFE N = 5,265	AS-IS N = 1,754
Sex			
Female	1,087 (51%)	2,595 (49%)	870 (50%)
Male	1,059 (49%)	2,670 (51%)	884 (50%)
SCD Subtype			
SCD-SS	1,945 (91%)	5,185 (98%)	1,707 (97%)
SCD-Sβ^0	71 (3%)	< 80 (< 2%)	47 (2.7%)
SCD Unspec.	131 (6%)	< 5 (< 1%)	0 (0%)
Years of Follow-Up[1]	5.5 (NR)	5.8 (5.5)	5.3 (4.8)

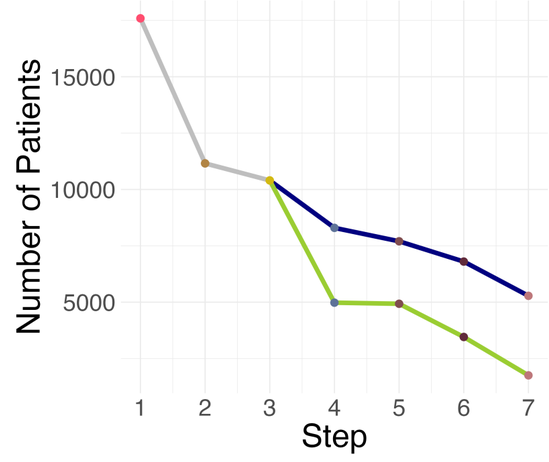
[1]Mean (Median)

Results (cont.)

Data Quality Remediation

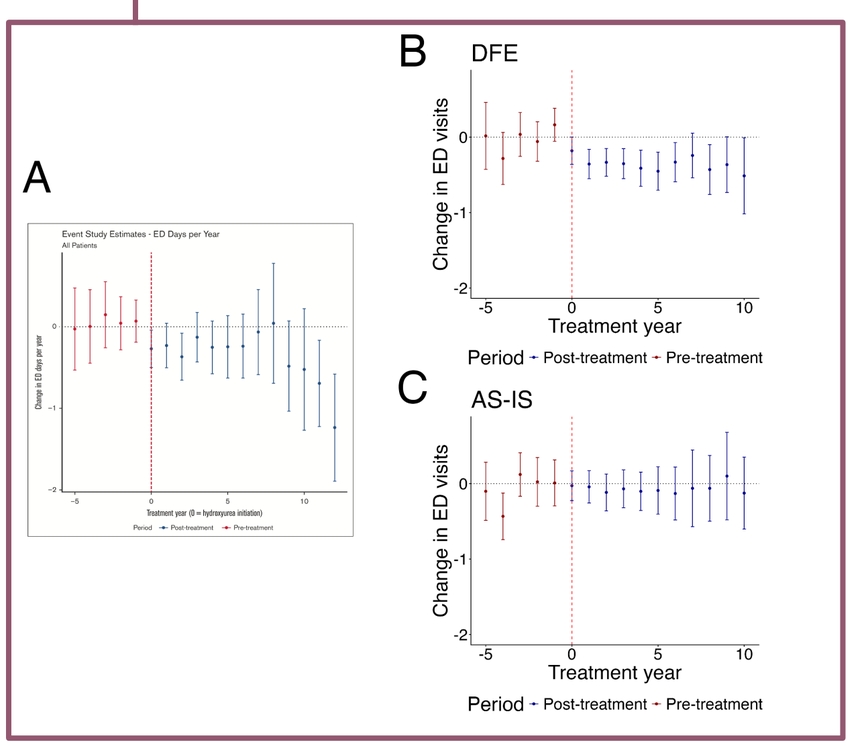


- Requiring 2+ SCD diagnoses instead of 1+ helped improve face validity of results related to utilization metrics & key study variables at several PEDSnet institutions.
- Iterative DQ evaluation helped to increase overall cohort size in the DFE cohort (blue) compared to the AS-IS cohort (green)



Difference-In-Difference Analysis

	ED Visits	Hospitalizations	Average HGB
(A) CHOA	-0.36 (CI -0.57,-0.16)	-0.84 (CI -1.51,-0.17)	0.56 (CI 0.39,0.73)
(B) DFE	-0.44 (CI -0.60,-0.26)	-1.93 (CI -2.46,-1.40)	0.10 (CI -0.03,0.22)
(C) AS-IS	-0.08 (CI -0.26,0.09)	-0.61 (CI -1.24,0.02)	-0.44 (CI -0.67,-0.22)



Conclusion

This replication study highlights the significant impact robust SSDQ analyses can have on analytic results. Despite all cohorts exhibiting similar descriptive characteristics, we only observed similar effects in the DFE and CHOA cohorts, while the AS-IS cohort exhibited attenuated effects and poor face validity. Without thorough evaluation of the source dataset and iterative refinement based on findings, results have the potential to lose power and validity. Additionally, the use of a standardized tool to support this level of rigorous evaluation facilitated systemic documentation and improved the transparency of study results. This ensures that results are more easily comparable across studies and supports future replication efforts.

These findings highlight the need for more comprehensive data fitness assessment and reporting in the research community. Especially as large EHR datasets become more openly accessible, investigators should approach each analysis with the appropriate level of methodological rigor to preserve the validity of results and support further replicability efforts.

References

1. Razzaghi H., Dickinson K., Wieand K., et al. (2026). A multifaceted approach to advancing data quality and fitness standards in multi-institutional networks. *J Am Med Inform Assoc, 33*(2), 371–382. https://doi.org/10.1093/jamia/ocaf181
2. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. (2021). Increasing trust in real-world evidence through evaluation of observational data quality. *J Am Med Inform Assoc, 28*(10):2251–2257. https://doi.org/10.1093/jamia/ocab132
3. Razzaghi, H., Wieand, K., Dickinson, K. L., Kahn, M. G., Roy, J., Blacketer, C., Christakis, D. A., Forrest, C. B., Greenberg, J., Lehmann, H. P., Marsolo, K. A., Sciolla, J., Weiner, M. G., Weiskopf, N. G., & Bailey, L. C. (2026). Beyond missingness: Systematizing methods for comprehensive data fitness assessment in clinical research. *Journal of Medical Internet Research*, *28*, e76398–e76398. https://doi.org/10.2196/76398
4. George, P., Kalmus, G., Lane, P. A., Lam, W., Lipscomb, J., & Howard, D. (2025). Evaluating the long-term benefits of hydroxyurea in pediatric sickle cell anemia. *Blood Advances*, *9*(14), 3585–3593. https://doi.org/10.1182/bloodadvances.2024015564
5. Razzaghi, H., Wieand, K., Pinkney, A., & Bailey, C. (2026). Systematic data fitness assessment improves validity and replicability of research using real-world data. *medRxiv*. https://doi.org/10.64898/2026.08.05.26359818

PEDSnet has been developed with funding from the Patient-Centered Outcomes Research Institute (PCORI); PEDSnet's participation in PCORnet is funded through PCORI award RI-CHOP-01-PS10. Support to develop the methods for this project was supported through PCORI Contract ME-2020C3-21199.

PEDSnet a pediatric learning health system