ARTICLE

Infrastructure

Every Day I Write the Book: Transforming EHR Data from Daily Practice into New Knowledge

Robert L. Phillips Jr., MD, MSPH and Andrew W. Bazemore, MD, MPH

Despite producing mountains of data in the daily course of care, the documentation labors of frontline clinicians currently return very little value to them or to the health system. The potential of these painstakingly collected data are enormous and clinical registries can extract the extraordinary capacity of these data and transform them into research-ready datasets while protecting the confidentiality of the patients and clinicians. Clinical registries represent transformative tools for primary care research, bringing together the dimensions of clinical practice, research, quality improvement, and policy impact from a large, nationally reflective, diverse sample of practices and patients. The PRIME Registry is one such clinical registry that extracts electronic health record data from more than 600 primary care practices across the United States that is helping advance research, improve quality, and shape the policies required to achieve high-performing primary care for all. Other examples of primary care registries exist, but most of the painstakingly captured data from frontline care remains fallow. Enabling use of these data are important for research, to prevent harm from mis-trained machine learning algorithms, and for monitoring the health of the public. (J Am Board Fam Med 2024;37:S138–S143.)

Introduction

More than 2 decades ago, Stange and Crabtree conducted groundbreaking observational studies in primary care clinics, novel for their capture of detailed elements of routine care and for characterizing the practices, their workforce, and their communities.[1,2] These studies produced dozens of secondary studies over a decade that lifted the veil on the details and relationships at the heart of primary care, where more than half a billion health care visits take place annually in the United States. More importantly they showed the richness and value of trusted health care relationships in the setting where more than 8 in 10 people in the US go when they have health concerns. At the time, primary care was an unexplored continent where National Institutes of Health (NIH) rarely funded expeditions or sought to establish outposts. Twenty years later it remains unexplored, but the stakes are now higher since the consequences are not just about ignorance, the advent of new technologies threaten to leave it behind or even do harm to the millions of people who come here for care.

The 2021 National Academies report *Implementing High Quality Primary Care* updated the considerable evidence that primary care remains the largest and most widely distributed platform of US Health care delivery and still devoid of sufficient research investment.[3] The report acknowledged that primary care

Corresponding author: Robert L. Phillips, MD, MSPH, 1016 16th St NW Suite 800 Washington, DC 20036 (E-mail: bphillips@theabfm.org).

receives 5 to 7 cents of every health care dollar for care delivery and less than 3 cents of every 10 dollars invested in health-related research.[3,4] The hundreds of millions of visits that happen in primary care annually generate terabytes of structured and unstructured data that is almost entirely fallow–not producing any residual net benefit to patients, public health, or science. It is a tragedy of epic scale. Any future success for primary care research in an era of big data and large language models will depend on access to authentic data from primary care. The Framingham Study taught us that community-level primary care data and its epidemiology can, over time, help us understand patients' risks of developing diseases and preventing them. The US could enjoy the fruits of a thousand Framinghams every year and with improved diversity and fit for the patients we care for. Instead, we face the risk of the development of machine learning derived algorithms designed for primary care but trained on data from nonprimary care settings. Past disease-oriented guidelines proved untrustworthy for primary care, but what is coming in the Artificial Intelligence (AI)/Machine Learning (ML) revolution in health care has the potential to be far more dangerous if directly applied in electronic health records (EHRs).

Primary Care Clinical Registries

Compared with claims data, the traditional sandbox for health services research, registries permit broader tracking of health indicators and patient characteristics and offer deeper insights into the rationale behind physician decision making. Their promise for research can be understood through several key lenses: the advancement of high-quality patient care, the facilitation of primary care clinical research, sentinel surveillance for population health and security, training large language models, and evidence capable of informing health policy. Recognizing the absence of real primary care data in a growing universe of registries, the American Board of Family Medicine (ABFM) in 2016 created the PRIME Registry, a Qualified Clinical Data Registry that extracts EHR data from more than 600 primary care practices across the United States.[5] Over its nearly 8 years in operation, PRIME has gathered data from more than 1200 practices caring for more than 8 million patients living in all 50 states (Table 1). Data elements include patient demographics, diagnoses, medication, test results, imaging reports, encounter-specific data, care planning

Table 1. American Family Cohort Patients by Race and Ethnicity (2021)

Race	# patients	%
Non-Hispanic White	4,588,980	75.2
Non-Hispanic Black	540,023	8.8
Hispanic	757,990	12.4
Asian	149,211	2.4
American Indian/Alaska Native	51,130	0.8
Native Hawaiian/Pacific Islander	16,632	0.3
Multiple races	2,113	0.03
Other race/not reported	1,660,672	

conversations, patient-reported outcomes, and some limited clinician-specific details. All data are collected during routine clinical care of patients with the related goals of supporting practice-specific quality improvement and quality reporting for payment. Five different small-area deprivation indices have also been applied to patients based on their home address. The American Family Cohort (AFC) is a research dataset derived from the PRIME Registry and there are several versions curated for different research use cases.[6] To better understand the functions and intentions of the PRIME Registry we recommend seeing our publications about its launch and the PRIME Registry website.[5,7] To better understand the American Family Cohort, the researchable data derivatives, partnerships, and publications, please see the AFC website.[6]

PRIME is hardly alone. The federal government has funded 9 clinical research networks through PCORNet.[8] Health Center Controlled Networks also enable tapping of EHRs of federally qualified health centers to create research capacity.[9] There is tremendous capacity for these real world, primary care data to be used to inform the epidemiology, care, and innovation but they are too often harnessed to disease specific research interests that have less utility for primary care. There are also countless other collections of EHR data from primary care captured by health systems that are simply siloed and untapped.

EHR registries collectively represent years of painstaking clinician data entry and conversations capturing care and planning; data capable of generating high value returns in the hands of purposeful primary care researchers. The current PRIME registry participants are estimated to write the equivalent of 10,000 "novels" (the average novel being 75,000-words long) each month in addition to a mountain of structured data. These large, longitudinal, frontline

epidemiologic, relationship-rich data are an underutilized research asset. The volume and duration of care they represent are not an unexplored continent, they are the health research equivalent of the earth's oceans or what lies below its mantle. Currently, we only study what washes up on shore or erupts on the surface. And it will become more important as we look to train machine learning and large language models to work for primary care.

Building Registry Research Capacity

Using EHR data for research is most often restricted by legal and regulatory concerns and by the lack of primary care research capacity prepared to manage enormous data sets. The ABFM has managed these problems in critical ways that could be enabling for others. Our partnership with the Stanford Center for Population Health Sciences has safely, legally lowered these barriers and streamlined the research process by offering a rich and research-ready national database for investigators. Protected Health Information from EHRs are managed under Business Associate Agreements (BAA) with practices and the vendor who pulls data from practices or receives on their behalf from their EHR vendor. Stanford also receives it under BAA and maintains it on secure servers with IRB oversight. Stanford cleans and normalizes the data (Observational Medical Outcomes Partnership (OMOP) transformation, for example) and adds additional enhancements (deprivation indices, for example) that improve value for research. They can then generate several derivatives based on risk reduction:

- **AFC OMOP DID:** De-Identified Data (DID) have been transformed into the OMOP common data model and are sufficient for most research projects.
- **AFC OMOP LDS:** Limited Data Set (LDS) data have additional variables not available in the DID data such as more granular date and geography variables. These data are suitable for studies which require overlay of social, environmental, or similar exposures.
- **AFC OMOP RIF:** Research Identifiable Files (RIF) data have had person readable identifiers removed but retain other information such granular dates, geography and detailed health data. These data are highly restricted and require significant additional regulatory and security steps to access.

- **AFC LDS and RIF including de-identified notes:** For research which requires data that have not been transformed to OMOP.

Researchers using registries like PRIME access large diverse samples difficult to compile independently, helping to generate reliable and broadly applicable results unavailable in single system data sources. This is particularly valuable for studying less common conditions or the effects of interventions across different populations and geographic locations. The diversity of the data in registries can also enhance comparative effectiveness research in primary care, offering a window on different health care interventions and informing best approaches for priority primary care conditions. The ABFM chose Stanford for the express purpose of allowing other, external partners to apply to use the data with appropriate review, protections, and in a reduced-risk data environment. Stanford also has many data scientists and related centers that are increasingly discovering the AFC dataset and are willing to partner and secure funding for questions of interest to primary care. Building primary care research capacity with AFC is also a goal. The ABFM has built up its own research capacity, its foundation is also investing to enhance capacity in family medicine departments, and ABFM data, like AFC, are attracting exceptional researchers to collaborate on topics important to primary care.[10–17]

Research using PRIME Registry has harmonized outcome measures for depression and explored the relationship between core dimensions of primary care (clinician and practice level Continuity, Comprehensiveness) and outcomes important to patients and policy makers.[18] It has furthered our understanding of gender-related income differences.[12] Research collaborations with the Centers for Disease Control and Prevention will produce nearly a dozen studies that offer an important window on pandemic response and recovery and could be the basis for longitudinal public health monitoring.[13,19,20] Research supported by National Library of Medicine enables a nationally respected AI/ML researcher to use AFC data to better characterize chronic kidney disease progression by race and ethnicity, and sets up potential to better understand prevention and treatment.[21] A grant from the Food and Drug Administration is exploring COVID-19 treatment patterns and equity.[22] A collaboration with the US Census Bureau will connect AFC data to many other, highly sensitive national datasets for the express

purpose of building more effective and reliable small-area social deprivation indices.[23]

Enabling Quality Improvement

Improving quality of primary care is not only a continuous process in health care, but an important dimension of primary care health services and implementation science research. For PRIME practices, their PRIME quality dashboard and patient care-gap tool offers benchmarks against national averages and Centers for Medicare & Medicaid Services (CMS) quality goals. Through these benchmarks, primary care physicians can identify areas where they excel and those that require improvement, enabling interventions to enhance patient care. Turning EHR data into quality measures also allows practices to report for a variety of quality payment programs with very low burden. PRIME's status as a Qualified Clinical Data Registry (certified by the Centers for Medicare and Medicaid Services) also gives it authority to propose, develop, test, and advance quality measures for certification. This now includes the Person-Centered Primary Care Measure, Continuity, Comprehensiveness, and Trust.[24] The ABFM considers these high-value primary care functions with growing evidence for their relationships with health, quality, and cost outcomes. PRIME is a laboratory for understanding which measures are important and making the case for advancing them so that primary care clinicians have incentives and support in achieving them.

The ABFM is partnering with the Arkansas State Department of Health, with Centers for Disease Control and Prevention (CDC) funding, to broaden PRIME adoption in the state with a primary purpose of monitoring blood pressure treatment and improvement. While the funding focus is very specific, we believe it will offer a much broader capacity to help practices identify care gaps and collaborate in improving them. Relatedly, we want the partnership to help the state and CDC to help practices through facilitation and new resources.

Sentinel Surveillance for Health Security and Population Health

Given its access to a broad sampling of the US population and its geographic and demographic heterogeneity, PRIME is well-equipped to enhance public health surveillance systems. Early in the COVID-19 pandemic, we differentiated patterns of Influenza Like Illness from COVID-19 and demonstrated their patterns across time and geography better than existing primary care disease monitoring programs at CDC.[12] Existing systems are currently ill-equipped to address emerging behavioral, chronic, and infectious threats to population health, manmade or epidemic and across subgroups of the US population. By combining regular clinical electronic data extraction and predictive modeling, a primary care registry like PRIME offers the potential for regular population health updates and threat monitoring. We are exploring federal interest in supporting the expansion of PRIME for this capacity.

Shaping Health Policy

Use of the PRIME Registry is governed by the Center for Professionalism & Value in Health Care which aims to use evidence to improve the health care environment for health care professionals and patients. The insights gained from the PRIME Registry are already demonstrating important policy opportunities. By providing a clearer picture of primary care effectiveness and the challenges faced by family physicians, PRIME is informing policy makers about the areas where resources are most needed. The aggregation of this data at a national level allows for a more nuanced understanding of health care delivery and outcomes across the country, which is essential for shaping policies that aim to improve health for all. One of the first publications out of the CDC collaboration made it abundantly clear that primary care was not part of the COVID-19 vaccination strategy, but more importantly, it also revealed that the federal vaccination registry that was created for the pandemic was not interoperable with most EHRs in use in primary care—practices were not only unable to give vaccine, they could not easily see which of their patients were vaccinated.[13] The ABFM believes that the data from PRIME can be used to support the case for primary care-focused reforms and help demonstrate the value of primary care services.

Building Out Primary Care Registry Capacity

The ABFM's experience with PRIME suggests that there is much more capacity for other primary care registries, like those in PCORNet and the Health

Center Control Networks. This includes research capacity in those networks, but also ways to safely enable other researchers to use them. There are also many other siloed primary care data within large health systems that could join the AFC or develop similar research capacity using its model. A confederation of these data holders might be appealing to federal health and research agencies to develop program support for research, AI/ML device development and testing, or health security monitoring. The PRIME Registry helps turn practice EHR data into information that they can use to track quality, identify care gaps, assess population health, and report quality to anyone they wish. Any practice interested in learning more or in joining PRIME should go to https://primeregisry.org/ where they can also sign up for a demonstration of the PRIME dashboard and the Population Health Assessment Engine (PHATE).[7] Guidance to researchers for assessing or accessing AFC data can be found at https://americanfamilycohort.org/.[6]

Conclusion

Stange and Crabtree provided an important window on primary care before EHRs.[1,2] While primary care data registries like PRIME lack the richness of those direct observation studies, they do offer a much larger and more longitudinal window on the care provided to millions of people every day. Registries represent transformative tools for primary care research, bringing together the dimensions of clinical practice, research, quality improvement, and policy impact from a large, nationally reflective, diverse sample of practices and patients. Despite producing mountains of data in the daily course of care, the documentation labors of frontline clinicians currently return very little value to them or to the health system. The potential of these painstakingly collected data are enormous but extracting the extraordinary capacity of these data will require support.

As health care delivery science continues to evolve toward more large data-driven approaches, registries and extraction tools capturing authentic primary care information like PRIME will likely become even more integral to the functioning of primary care. By leveraging the potential of the PRIME Registry, the primary care community can continue to advance research, improve quality, and shape the policies required to achieve high-performing primary care for all.

To see this article online, please go to: http://jabfm.org/content/37/S2/S148.full.

References

1. Crabtree BF, Miller WL, Tallia AF, et al. Delivery of clinical preventive services in family medicine offices. Ann Fam Med 2005;3:430–5.
2. Stange KC, Zyzanski SJ, Jaén CR, et al. Illuminating the "black box": a description of 4454 patient visits to 138 family physicians. J Fam Pract 1998;46:377–89.
3. National Academies of Sciences, Engineering, and Medicine. *Implementing High-Quality Primary Care: Rebuilding the Foundation of Health Care*. The National Academies Press; 2021.
4. Mendel P, Gidengil CA, Tomoaia-Cotisel A, et al. *Health Services and Primary Care Research Study: Comprehensive Report*. RAND Corporation; 2020.
5. Phillips R. The PRIME Registry helps thousands of primary care clinicians liberate EHR data and prepare for MIPS. J Am Board Fam Med 2017;30:559–559.
6. The American Board of Family Medicine. The American family cohort. Accessed December 17, 2023. Available at: https://americanfamilycohort.org/.
7. Primecreativeteam. About Prime Registry. Prime Registry. Accessed January 5, 2024. Available at: https://primeregistry.org/.
8. PCORNet. The National Patient-Centered Clinical Research Network. Accessed December 17, 2023. Available at: https://pcornet.org/network/.
9. Health Resources and Services Administration. Health center controlled networks | Bureau of Primary Health Care. Accessed December 17, 2023. Available at: https://bphc.hrsa.gov/technical-assistance/strategic-partnerships/health-center-controlled-networks.
10. American Board of Family Medicine. ABFM Foundation selects inaugural cohort of AI/ML grant recipients | ABFM | American Board of Family Medicine. Publications, News, and Press Releases. Published June 29, 2022. Accessed March 19, 2024. Available at: https://www.theabfm.org/communications/news/abfm-foundation-selects-inaugural-cohort-aiml-grant-recipients.
11. Dai M, Peterson LE, Phillips RL. Quality changes among primary care clinicians participating in the transforming clinical practice initiative. J Healthc Qual 2021;43:e64–e69.

12. Ganguli I, Mulligan KL, Phillips RL, Basu S. How the gender wage gap for primary care physicians differs by compensation approach: a microsimulation study. Ann Intern Med 2022;175:1135–42.

13. Hao S, Rehkopf DH, Velasquez E, Vala A, Bazemore AW, Phillips RL. COVID-19 vaccine strategy left small primary care practices on the sidelines. Health Aff (Millwood) 2023;42:1147–51.

14. Hendrix N, Phillips RL, Bazemore AW. How do family physicians document patients' social needs in electronic health records? J Am Board Fam Med 2023;36:510–2.

15. Hendrix N, Bazemore A, Holmgren AJ, et al. Variation in family physicians' experiences across different electronic health record platforms: a descriptive study. J Gen Intern Med 2023;38:2980–7. Published online March 23.

16. Casalino LP, Li J, Peterson LE, et al. Relationship between physician burnout and the quality and cost of care for Medicare beneficiaries is complex. Health Aff (Millwood) 2022;41:549–56.

17. Rittenhouse DR, Wiley JA, Peterson LE, Casalino LP, Phillips RL. Meaningful use and medical home functionality in primary care practice. Health Aff (Millwood) 2020;39:1977–83.

18. Leavy MB, Boussios C, Phillips RL Jr, et al. Calculation of harmonized depression outcome measures in registries. In: *Outcome Measure Harmonization and Data Infrastructure for Patient-Centered Outcomes Research in Depression: Final Report [Internet]*. Agency for Healthcare Research and Quality (US); 2022.

19. Phillips RL, Rehkopf DH, Vala A, Bazemore AW, Peterson L, Chu I. Clinical registries could improve influenza like illness and COVID-19 surveillance. COVID-19 Collection: Annals of Family Medicine. Available at: http://hdl.handle.net/2027.42/154853.

20. Centers for Disease Control and Prevention. Outpatient electronic health record data from primary care and family medicine practices to support CDC's COVID-19 emergency response (American Board of Family Medicine). USASpending.gov. Published August 24, 2022. Accessed January 5, 2024. Available at: https://usaspending.gov/award/CONT_AWD_75D30122P12974_7523_-NONE-_-NONE-.

21. Rose S. *Novel Algorithmic Fairness Tools for Reducing Health Disparities in Primary Care*. Stanford University; 2023. Accessed January 5, 2024. Available at: https://reporter.nih.gov/search/mqu2r0TXTkGzyTdSuQK8kg/project-details/10676234.

22. Rehkopf DH. *Assessing the Racial, Ethnic, and Geographic Equitability of COVID-19 Treatments in a Primary Care Population*. Stanford University; 2023. Accessed January 5, 2024. Available at: https://reporter.nih.gov/search/jaOBlZtTmEizi-a3a4zjtw/projects.

23. Bureau UC. Partnership: American Board of Family Medicine and Stanford University. Census.gov. Accessed January 5, 2024. Available at: https://www.census.gov/programs-surveys/ehealth/information/abfm-partnership.html.

24. The Center for Professionalism & Value in Health Care. Measures that matter. Published 2021. Accessed September 29, 2021. Available at: https://y7i6a8f8.rocketcdn.me/wp-content/uploads/2020/09/Measures-That-Matter-Flyer-2020.pdf.