This online first version has been peer-reviewed, accepted and edited, but not formatted and finalized with corrections from authors and proofreaders

Health system evaluation: new options, opportunities and limits

Kevin Croke,[a] Edwine Barasa[b] & Margaret E Kruk[a]

[a] Harvard TH Chan School of Public Health, 677 Huntington Ave, Boston, MA 02459, United States of America.

[b] Kenya Medical Research Institute/Wellcome Trust Research Programme, Nairobi, Kenya.

Correspondence to Kevin Croke (email: kcroke@hsph.harvard.edu).

(Submitted: 20 January 2023 – Revised version received: 1 March 2024 – Accepted: 2 March 2024 – Published online: 28 May 2024)

High-quality evaluation is critical for health systems because it enables the best use of scarce resources and helps policy-makers learn what works best in their setting. What works cannot be assumed: since health systems are complex, system reforms have long causal chains and multiple interacting components. Many promising health system interventions, even those that increase intervention coverage, fail to improve health outcomes such as mortality and morbidity. However, many health system reforms, especially in low- and middle-income countries, are never evaluated due to data limitations and scarce resources for health systems research.

Counterfactual-based designs (that is, evaluations in which research design enables inference about the causal impact of a policy) are challenging to implement for health system reforms. Reforms are often applied to the whole system, leaving no obvious control group, or are assigned to high-level administrative units, limiting the number of treated and comparison units. Reforms may be targeted to areas or groups for political reasons, limiting ability to randomize and constraining generalizability. Many researchers have seen these challenges as reasons why rigorous evaluations of complex health system reforms are unlikely to succeed, promoting instead so-called realist designs rooted in largely qualitative methods. However, in recent years, innovations in evaluation methods, approach and data have opened new possibilities for health system evaluation. These innovations include large-scale randomized health system trials, causal inference methods for better non-randomized inference and new technologies for data collection and analysis, including big data. These applications, which have emerged from disparate academic disciplines and from practice, may not be fully appreciated by applied health systems researchers.

A first innovation is the growing application of randomized controlled trials to system questions. Randomization has often been considered infeasible for health system questions. For example, a study shows that 79% (139/176) of intervention evaluation papers in top medical, economics, and health services journals in the United States of America were randomized controlled trials, compared to fewer than one fifth of health delivery (that is, health system) papers.[1] However, use of randomized designs for health system evaluation in low- and middle-income countries is increasing. Recent examples include evaluations of performance-based financing in Nigeria,[2] subsidized health insurance in Indonesia,[3] community health worker recruitment and supply chain organization for medicine delivery in Zambia[4,5] and point-of-care quality interventions in northern India.[6] Beyond maximizing internal validity, randomized controlled trials also allow researchers to test causal mechanisms, including predictions derived from theory. Direct tests of theory can enable systematic, linked accumulation of knowledge on important questions. These studies have had important implications for practice, for example by limiting enthusiasm about the potential of performance-based financing or coaching interventions to improve quality of care. In the absence of high-quality randomized evidence, advocates on opposing sides might have continued to cite competing non-randomized studies. Randomization can be difficult for political and practical reasons. Yet these proofs by existence demonstrate that large-scale, system-level randomized controlled trials should not be considered impossible beforehand, particularly when governments and/or sponsors are keen to learn and engaged early in the process.

When randomization is not possible, rigorous evaluations of health system reforms with careful attention to counterfactual comparison has become increasingly feasible using methods such as difference-in-difference or regression discontinuity designs. Application of these methods has been hindered in the past by limited data availability. Yet the data picture has changed for the better in many low- and middle-income countries. Household surveys such as the Demographic and Health Surveys have expanded their topical and geographic coverage, and are now routinely geocoded. When multiple national survey rounds take place before and after programme scale-up, difference-in-difference research designs can often be used to estimate the impact of policy. Administrative data from vital registration systems, national health management information systems or national insurance programme claims data, long used in high-income countries, are increasingly usable for such research in middle-income countries. Brazil provides several examples: one study uses the staggered expansion of Brazil's *Programa Saúde da Família* in a difference-in-difference framework to demonstrate that the programme substantially reduced infant and maternal mortality.[7] By contrast, and using similar data and methods, another study shows that the *Mais Medicos* programme, in which expatriate doctors were deployed to underserved communities in Brazil, did not affect infant mortality.[8]

This approach has been more limited in regions that lack comprehensive vital registration. In these settings, administrative databased evaluations are often limited to utilization data, aggregated on the platform of the health management information system DHIS2. These data systems have faced challenges with data quality as well as completeness. Health management information system data capture what happens in facilities, missing outcomes (including mortality) that occur at home, and typically do not capture individual-level data. Yet even with these limitations, these data have been increasingly leveraged in studies for which service utilization is the primary outcome. For example, these data have been used to demonstrate the impact of the coronavirus disease 2019 (COVID-19) pandemic on health system utilization.[9]

These challenges of data completeness and quality have generated enthusiasm about potential uses of technology, including big data, to enable more health system evaluation, including in settings with limited administrative data. Experience so far demonstrates promise but also grounds for caution. On the positive side, digital technologies have been used to improve surveys, digitize routine data collection, expand demographic surveillance systems and integrate remotely sensed data. Mass mobile phone ownership has opened new possibilities for mobile data collection in low- and middle-income countries: during COVID-19, many researchers and institutions successfully implemented mobile phone data collection protocols. For example, recent multicountry mobile phone surveys have effectively captured health system performance data in low-income countries.[10] Digital technologies have also enabled expansion of health and demographic sentinel sites to nationally representative scale in some settings, such as Mozambique's Countrywide Mortality Surveillance for Action system. New digital platforms, such as the Socioeconomic High-resolution Rural-Urban Geographic Platform for India,[11] can now aggregate surveys, geospatial data and geographically coded administrative data.

Larger scale applications of big data (beyond survey and administrative data) have been creatively leveraged for some forms of health research in low- and middle-income countries. Mobile phone call data records have been used to study population movement, informing disease transmission dynamics. Social media posts have been used to predict disease outbreaks. Researchers have envisioned a future in which passively collected health status measures from wearable devices or health utilization from facility-based sensors, via the Internet of Things, can be used for evaluation. In development economics, researchers increasingly benefit from the fact that variables of interest such as night-time luminosity, housing infrastructure, temperature, pollution or land use are now observed at high frequency and resolution by satellites or other remote sensing apparatuses. Researchers train machine learning algorithms to measure poverty based on these observations, opening scope for new forms of economic and social policy evaluation.

Yet application of these data sources to health system policy evaluations in low- and middle-income countries is still nascent. Many public health applications of big data are used for mapping or prediction, which is extremely useful but distinct from policy evaluation. Key health outcomes of interest such as service utilization, health status, financial risk protection or population attitudes cannot be measured by satellites or other remote sensing tools. The same incomplete electrification and digitization of health facilities that currently limits evaluation designs using DHIS2 are likely to render big data from wearables and facility-based sensors unreliable for national-scale health system evaluation studies. New technologies can generate data with greater temporal and spatial coverage. However, they do not solve the health system evaluation problem because they create a situation in which data is relatively plentiful, but credible research designs and researchers are, in relative terms, scarce. New data sources open promising areas for health systems evaluation, but they cannot substitute for careful research design. Effective health system inquiry requires robust theoretical frameworks, supported by improvements in underlying administrative and vital registration data systems.

New approaches must also be rooted in an appreciation of the practicalities of policy change. Evaluation strategies must be compatible with a plausible theory of how organizations (including governments) learn and how they make policy decisions.[12] Health policy rarely changes based only on evidence. Policy evolves over time as competing coalitions of experts, politicians and stakeholders push for their preferred solutions. Evidence is only one input into this mix, along with values, public opinion, interest group pressure, previous studies and ideological and intellectual predispositions of policy-makers. Policy-makers' willingness to change their minds based on new evidence may be as much a function of the strength of their relationship and the depth of their trust with researchers, as it is with the technical rigor of the evidence.

Together with improved data and evaluation methods, investment is needed in the institutions that implement high-quality evaluations and participate in ongoing policy dialogues about the future of the health system. These institutions comprise the health research and policy community in low- and middle-income countries, including academia, think tanks, research units embedded in ministries and evidence-oriented nongovernmental organizations. These organizations are also the natural constituency to press governments to invest their own resources in better statistical systems, including both routine and survey data, which will enable better evaluation on an ongoing basis. Institutionalization of this process is not just a key ingredient in policy translation: it can also help strengthen the sustainability of evaluation and health system learning over time.

Competing interests:

None declared.

References

1. Finkelstein A, Taubman S. Randomize evaluations to improve health care delivery. Science. 2015 Feb 13;347(6223):720–2. https://doi.org/10.1126/science.aaa2362 PMID:25678649
2. Khanna M, Loevinsohn B, Pradhan E, Fadeyibi O, McGee K, Odutolu O, et al. Decentralized facility financing versus performance-based payments in primary health care: a large-scale randomized controlled trial in Nigeria. BMC Med. 2021 Sep 21;19(1):224. https://doi.org/10.1186/s12916-021-02092-4 PMID:34544415
3. Banerjee A, Finkelstein A, Hanna R, Olken BA, Ornaghi A, Sumarto S. The challenges of universal health insurance in developing countries: experimental evidence from Indonesia's national health insurance. Am Econ Rev. 2021 Sep 1;111(9):3035–63. https://doi.org/10.1257/aer.20200523
4. Vledder M, Friedman J, Sjöblom M, Brown T, Yadav P. Improving supply chain for essential drugs in low-income countries: results from a large scale randomized experiment in Zambia. Health Syst Reform. 2019;5(2):158–77. https://doi.org/10.1080/23288604.2019.1596050 PMID:31194645

5. Ashraf N, Bandiera O, Davenport E, Lee SS. Losing prosociality in the quest for talent? Sorting, selection, and productivity in the delivery of public services. Am Econ Rev. 2020 May;110(5):1355–94. https://doi.org/10.1257/aer.20180326

6. Semrau KEA, Hirschhorn LR, Marx Delaney M, Singh VP, Saurastri R, Sharma N, et al.; BetterBirth Trial Group. Outcomes of a coaching-based WHO safe childbirth checklist program in India. N Engl J Med. 2017 Dec 14;377(24):2313–24. https://doi.org/10.1056/NEJMoa1701075 PMID:29236628

7. Bhalotra SR, Rocha R, Soares RR. Does universalization of health work? Evidence from health systems restructuring and expansion in Brazil. Bonn: Institute of Labor Economics; 2019. Available from: https://papers.ssrn.com/abstract=3390099 [cited 2023 Apr 27].

8. Carrillo B, Feres J. Provider supply, utilization, and infant health: evidence from a physician distribution policy. Am Econ J Econ Policy. 2019;11(3):156–96. https://doi.org/10.1257/pol.20170619

9. Arsenault C, Gage A, Kim MK, Kapoor NR, Akweongo P, Amponsah F, et al. COVID-19 and resilience of healthcare systems in ten countries. Nat Med. 2022 Jun;28(6):1314–24. https://doi.org/10.1038/s41591-022-01750-1 PMID:35288697

10. Kruk ME, Kapoor NR, Lewis TP, Arsenault C, Boutsikari EC, Breda J, et al. Population confidence in the health system in 15 countries: results from the first round of the People’s Voice Survey. Lancet Glob Health. 2024 Jan;12(1):e100–11. https://doi.org/10.1016/S2214-109X(23)00499-0 PMID:38096882

11. Asher S, Lunt T, Matsuura R, Novosad P. Development research at high geographic resolution: an analysis of night-lights, firms, and poverty in India using the SHRUG open data platform. World Bank Econ Rev. 2021 Nov 1;35(4):845–71. https://doi.org/10.1093/wber/lhab003

12. Pritchett L. The politics of learning: directions for future research (RISE Working Paper). Oxford: RISE; 2018. Available from: https://riseprogramme.org/publications/politics-learning-directions-future-research [cited 2024 Feb 9].