there were insufficient treatment options, against the risks, including the observed mortality imbalance. The risk associated with inadequate treatment of tuberculosis includes the likely progression of disease, which would be fatal in some cases, and the development of increased antimycobacterial resistance not only for the patient, but also for broader populations at risk for acquiring tuberculosis. The limited indication of use for bedaquiline identifies a patient population for which there is considerable unmet need and a positive benefit–risk balance.[1] It is crucial that physicians and patients with multidrug-resistant tuberculosis carefully consider this information as well as the potential ramifications of inadequate treatment and increasing resistance.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

From the Office of Antimicrobial Products, Office of New Drugs, Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD.

1. Sirturo (bedaquiline) product insert. Silver Spring, MD: Food and Drug Administration (http://www.accessdata.fda.gov/drugsatfda_docs/label/2012/204384s000lbl.pdf).
2. Avorn J. Approval of a tuberculosis drug based on a paradoxical surrogate measure. JAMA 2013;309:1349-50.
3. Global tuberculosis report 2013. Geneva: World Health Organization, 2013 (http://apps.who.int/iris/bitstream/10665/91355/1/9789241564656_eng.pdf).
4. Tiemersma EW, van der Werf MJ, Borgdorff MW, Williams BG, Nagelkerke NJ. Natural history of tuberculosis: duration and fatality of untreated pulmonary tuberculosis in HIV negative patients: a systematic review. PLoS One 2011;6(4):e17601.
5. Center for Drug Evaluation and Research. Medical officer review: Sirturo (bedaquiline). Silver Spring, MD: Food and Drug Administration (http://www.accessdata.fda.gov/drugsatfda_docs/nda/2012/204384Orig1s000MedR_.pdf).

DOI: 10.1056/NEJMp1314385


# Did Hospital Engagement Networks Actually Improve Care?

Peter Pronovost, M.D., Ph.D., and Ashish K. Jha, M.D., M.P.H.

Everyone with a role in health care wants to improve the quality and safety of our delivery system. Recently, the Centers for Medicare and Medicaid Services (CMS) released results of its Partnership for Patients Program (PPP) and celebrated large improvements in patient outcomes.[1] But the PPP's weak study design and methods, combined with a lack of transparency and rigor in evaluation, make it difficult to determine whether the program improved care. Such deficiencies result in a failure to learn from improvement efforts and stifle progress toward a safer, more effective health care system.

CMS launched the PPP in December 2011 as a collaborative comprising 26 "hospital engagement networks" (HENs) representing more than 3700 hospitals, in an effort to reduce the rates of 10 types of harms and readmissions. The HENs work to identify and disseminate effective quality-improvement and patient-safety initiatives by developing learning collaboratives for their member facilities, and they direct training programs to teach hospitals how to improve patient safety. In a February 2013 webcast, CMS announced that the rates of early elective deliveries had dropped 48% among 681 hospitals in 20 HENs and that the national rate of all-cause readmissions had decreased from 19% to 17.8%, though it is unclear which HENs were included for each measure and what time periods were the pre- and post-intervention periods.[1]

These numbers appear impressive, but given the publicly available data and the approach CMS used, it's nearly impossible to tell whether the PPP actually led to better care. Three problems with the agency's evaluation and reporting of results raise concerns about the validity of its inferences: a weak design, a lack of valid metrics, and a lack of external peer review for its evaluation. Though the evaluation of many other CMS programs also lacks this basic level of rigor, given the large public investment in the PPP, estimated at $1 billion, and the strong public inferences about its impact, the lack of valid information about its effects is particularly troubling.

The design of a quality-improvement program influences our ability to make reasonable inferences about its benefits to patients. Although individual HENs may have used more rigorous methods, the overall PPP evaluation had three important weaknesses: it used a pre–post design with only single points in the pre and post periods, did not have concurrent controls, and did not specify the pre and post periods a priori. Such an approach is highly subject to bias.[2] Several recent examples suggest that some patient-safety interventions appear to lead

to improvements but are no more effective than controls. For example, a 2011 evaluation of a multiple-component patient-safety intervention in the United Kingdom showed improvement in control hospitals that was as robust as that in intervention hospitals.[3] Without appropriate controls, it is difficult to know whether the measured effects of an intervention actually reflect secular trends.

There are alternatives available, including a randomized or even a cluster-randomized trial. If such trials were not feasible, CMS could have used other robust design approaches, such as an interrupted time-series study with concurrent controls. Rather than having a single pre time period and a single post time period, this design entails repeated measurements of the safety indicators before and after the intervention in both HEN and non-HEN hospitals. Such an approach would have provided more valid inferences about the effects of the program, with few additional costs.

Beyond using a poor design, CMS did not use standardized and validated performance measures across all participating hospitals — further hampering inferences about the program's effects. To support engagement, CMS allowed each HEN to define its own performance measures, with little focus on data quality control. For example, in reporting catheter-associated urinary tract infections, HENs could rely on the Centers for Disease Control and Prevention (CDC) definitions, an approach requiring clinical data, or they could use administrative data. But administrative data on these infections are widely considered insensitive and are subject to variation and changes in coding practices. Furthermore, since variation in measures rendered it impossible to compare all HEN hospitals, the ability to evaluate "improvement" was limited. For example, CMS suggested that HENs use the National Database of Nursing Quality Indicators definitions to evaluate pressure ulcers. Most HENS did not use these definitions, and most of the data on pressure-ulcer improvements came from a minority of institutions.

CMS also required HENs and participating hospitals to submit a large number of process measures of unknown validity. It is essential to use validated measures — ideally those endorsed by the National Quality Forum — unless there is a compelling reason not to. Large-scale quality-improvement efforts are most successful when they include standardized performance measures that clinicians believe are valid, when clinicians receive rapid feedback on performance, and when clinicians are encouraged to modify the intervention to fit their local context.[4] In instances where validated measures are unavailable, instead of using poor quality metrics, CMS can have an agency such as the Agency for Healthcare Research and Quality (AHRQ) or the CDC develop measures rapidly.

Finally, CMS made — and presented publicly — inferences about its program's benefits without having subjected its work to independent evaluation or peer review. Peer review, though imperfect, is a powerful quality control. Indeed, the peer-review process would probably have raised many of the concerns highlighted here. Such a review might have prompted CMS to change its evaluation plan or at least provide substantially more data than it has provided to date.

The PPP involved an investment of nearly $1 billion to improve care — three times the annual budget of the AHRQ, the lead federal funding agency for implementation science, which often lacks resources for promising projects. With such a sizable investment, CMS could have supported a better evaluation. It could have randomized HENs or hospitals to receive interventions earlier or later; used standardized, validated measures across the HENs; built in basic data quality controls; and independently collected qualitative information alongside quantitative data to learn not just whether the interventions worked but also how and why they did, thereby advancing our understanding of the mechanisms and context of improvement science. These changes would have allowed the country to learn so much more.

The lack of a careful evaluation is symptomatic of a broader problem: some members of the quality-improvement community eschew even modestly rigorous methods, believing that one can simply "know" if an intervention worked.[5] Though maintaining hope and optimism among clinicians is important, when untested interventions are implemented widely, they often fail to improve care. The confidence we can have in an intervention's efficacy is directly related to the rigor with which it is designed, implemented, and evaluated. Given the strong desire to improve care and the conflicts of interest we all face in evaluating our own work, subjecting all evaluations to external examination is critical.

The field of improvement science is still in its infancy. Given the magnitude of the quality and

cost problems in health care and the amount of money invested in mitigating these problems, the public, providers, and policymakers need to have confidence that money used to improve care is being well spent. It's true that improvement science requires mixed methods and is difficult, but all good science is difficult. Failing to attend closely to issues of design, methods, and metrics leaves us with little confidence in an intervention. For the PPP, which required thousands of hours of clinicians' time and large sums of money, that lack of confidence is particularly unfortunate. More important, the failure to generate valid, reliable information hampers our ability to improve future interventions, because we are no closer to understanding how to improve care than we were before the PPP. And that is the biggest cost of all.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org

From the Armstrong Institute for Patient Safety and Quality, Johns Hopkins Medicine, and the Department of Anesthesiology and Critical Care Medicine, the Department of Surgery, and the Department of Health Policy and Management, Johns Hopkins University — both in Baltimore (P.P.): and the Department of Health Policy and Management, Harvard School of Public Health, the Department of Medicine, Division of General Internal Medicine, Brigham and Women's Hospital, and the Veterans Affairs Boston Healthcare System — all in Boston (A.K.J.).

**1.** McKinney M. Federal push shows gains: quality campaign makes headway, but lack of standardized data blurs picture. Mod Healthc 2013;43:8-9.
**2.** Fan E, Laupacis A, Pronovost PJ, Guyatt GH, Needham DM. How to use an article about quality improvement. JAMA 2010;304:2279-87.
**3.** Benning A, Dixon-Woods M, Nwulu U, et al. Multiple component patient safety intervention in English hospitals: controlled evaluation of second phase. BMJ 2011;342:d199.
**4.** Dixon-Woods M, Bosk CL, Aveling EL, Goeschel CA, Pronovost PJ. Explaining Michigan: developing an ex post theory of a quality improvement program. Milbank Q 2011;89:167-205.
**5.** Auerbach AD, Landefeld CS, Shojania KG. The tension between needing to improve care and knowing how to do it. N Engl J Med 2007;357:608-13.

DOI: 10.1056/NEJMp1405800


# The Impact and Evolution of Medicare Part D

Julie M. Donohue, Ph.D.

It has been 10 years since the Medicare Prescription Drug, Improvement, and Modernization Act was signed by President George W. Bush, and 8 years since its centerpiece — a new Medicare drug benefit (Part D) — was implemented. Criticisms during Part D's implementation — citing poor communication with beneficiaries, computer glitches, complicated plan choices, and cost concerns — bear a striking resemblance to those currently voiced about the Affordable Care Act (ACA). Yet Medicare Part D successfully expanded drug benefits to millions of beneficiaries and improved access to medications, at lower-than-expected cost.

Part D has its challenges, however, and policymakers continue to modify various aspects of the program. The concerns raised about Part D relate to the key choices policymakers face when establishing any new insurance program — regarding enrollment, competition, coverage, and pricing.

The first question was whether Medicare beneficiaries would enroll. Unlike the ACA, Part D was established as a voluntary benefit. That decision raised concerns that too few people would participate and that enrollees would be sicker than average, which would lead to higher premiums and even lower enrollment in subsequent years. The legislation therefore included a late-enrollment penalty, although surveys suggest that few beneficiaries were aware of it.

In fact, Part D participation has been high. Kaiser Family Foundation data indicate that by June 2006 (8 months after enrollment began), Part D covered 22.5 million beneficiaries (53% of Medicare beneficiaries). Enrollment grew to 35.7 million beneficiaries (69%) in 2013. Another 20% of beneficiaries have coverage through other sources (e.g., retiree health plans). Thus, 10% still lack drug coverage — somewhat more than originally forecast (see graph). ACA participation may be higher because of the mandate that individuals obtain coverage.

The second question was whether beneficiaries would have enough plan choice and would make good choices. Part D established a new insurance product, inviting plans to compete for enrollees in 34 regions. Competition was intended to lower premiums and allow beneficiaries to find plans that would best meet their needs. Concerns centered on whether enough plans would