JAMA Health Forum™

JAMA Forum

Overuse and Underuse of Health Care: New Insights From Economics and Machine Learning

Katherine Baicker, PhD; Ziad Obermeyer, MD

Author affiliations and article information are listed at the end of this article.

Rampant overuse of health care in the US is plain to see. The traditional explanation from health economics is straightforward: bad incentives. Paying physicians too much leads to more care—whether patients need it or not. But paying physicians too little risks limiting patients' access to care. There is a similar trade-off in patient cost-sharing. This suggests that health policy, from insurance design to payment models, should calibrate payments to be high enough to discourage stinting and promote access, but low enough to avoid overuse of low-value care that threatens affordability and financial sustainability.

But mounting evidence suggests this trade-off is not so simple. Across vastly different settings and incentive schemes, there is both overuse and underuse at the same time. The traditional lever of dialing payments up or down is thus likely to mitigate one problem while exacerbating the other. Instead, new tools are required to help physicians do better, rather than do more or do less across the board.

Emerging research blending behavioral economics with cutting-edge data analytics paints a vivid picture of coexisting overuse and underuse—and suggests novel ways to counter both. For example, in a recent study, we built a machine learning algorithm to study how physicians test for acute coronary syndrome (ACS) in the emergency department.[1] Algorithms are well suited to this task: they can predict which patients have such low likelihood of testing positive that it is not worth doing the test and which patients have such high risk of ACS that swift diagnosis and treatment can be lifesaving. Knowing which patient is which, drawing on rich data available at decision time, can help focus testing on patients who are most likely to benefit.

Comparing algorithmic predictions to physicians' decisions reveals substantial overtesting.[1] About two-thirds of tests were performed on patients with predictably low risk, making the tests extremely low value—some costing up to $1 million per life-year saved. But, critically, we also find substantial undertesting, with predictably high-risk patients going untested and then having adverse outcomes of missed ACS, including death. These findings suggest that reallocating low-value tests to high-risk untested patients could save lives, at a cost of only $46 017 per life-year.

These numbers point to potential large-scale inefficiency, but more direct evidence comes from a "natural experiment" leveraging the fact that different teams of clinicians order tests at higher or lower rates. Such widespread variation in care practices, occurring even within a given hospital from 1 emergency department shift to the next, means the likelihood of a given patient getting tested depends on the idiosyncratic arrival time at the emergency department. This design, increasingly popular in health research because it simulates a randomized trial,[2,3] lets us measure the effect of testing on health outcomes.

If we simply ask whether there is too much or too little testing, the answer is clear: on average, patient outcomes are no better on high-testing vs low-testing shifts—we would be better off cutting those extra tests. This average effect, though, lumps the handful of high-risk patients together with relatively low-risk patients. Fine-grained machine learning predictions can tell these apart—and tell a different story. Low-testing shifts appear efficient on average, but cut back on testing for both high-risk patients (for whom tests have large benefits) and low-risk patients (for whom tests have little or no benefit). High-testing shifts similarly test both high-risk and low-risk patients more.

So it is no surprise that higher across-the-board testing has little aggregate health benefit because most patients are low risk. But for the small fraction of predictably high-risk patients, we find a dramatic reduction in adverse events and death—34% lower 1-year mortality—when they arrive during higher-testing shifts.[1] We estimate that the optimal policy would cut testing by 46.8% overall, but with a 62.4% reduction in the tests physicians currently do and a 15.6% increase in testing for patients who currently go untested.

Importantly, these suboptimal testing patterns exist in both top-ranked academic medical centers and nationally representative samples. A study of computed tomography pulmonary angiography in national Medicare emergency department visits demonstrates similarly large-scale overuse and underuse.[4] Another study shows how variation in radiologists' diagnostic skill can drive both overdiagnosis and underdiagnosis of pneumonia.[5] Yet another study shows that hospitals' relative level of specialization in invasive vs medical management of ACS can give rise to what amounts to both overuse and underuse.[6]

The simultaneous presence of overuse and underuse poses a challenge for traditional incentive-based economic models: physicians prescribe more care when they are paid more for it, and patients use more care when they pay less for it. Although this can explain overuse, it cannot explain widespread underuse of care that is both high-value and highly reimbursed. An emerging literature in behavioral economics illustrates the important role of nonfinancial as well as financial factors in driving care decisions, a phenomenon referred to as "behavioral hazard" (vs the core economics principle of "moral hazard" driven by misaligned financial incentives).[7]

Some behavioral factors, such as physicians' skill, specialization, or intrinsic motivation, can enhance performance.[5,6,8] Others, such as cognitive limitations and biases, can worsen performance. We found that physicians' ACS testing decisions rely on an overly simple risk model that focuses on a handful of variables closely linked to ACS risk but neglects hundreds of others—ones that a machine learning model can capture. Physicians also put too much weight on salient variables, such as chest pain, relative to those variables' actual predictive value. This complexity illustrates the value of machine learning–based decision aids, which can be used to construct risk scores that capture the richness of individual patients' histories to help physicians deploy tests more effectively.

The combination of behavioral economics and machine learning can not only generate new insights into observed patterns of care, but also inform redesign of both payments and decision aids to better target care in real time. We believe this approach holds great promise for reducing waste while improving outcomes. But rigorous real-world testing is crucial before widespread adoption, and we are currently undertaking a randomized trial for the algorithm we developed.

Rigorous evaluation is important because these new tools, for all their promise, have major limitations. Like other statistical approaches, algorithms can only learn from the data they have. These data are inherently incomplete because they come from the existing health system and existing physician decision-making patterns. For example, our natural experiment[1] indicates that increasing testing from the lowest rate we observed (18.1%) to the highest (32.3%) would reduce mortality, but we cannot recommend testing all high-risk patients, because we do not see what would have happened at even higher rates.

This is not an abstract statistical point: physicians may leave many high-risk patients untested because they have access to important information that the algorithm does not, including the physical examination findings (eg, the bruise on a high-risk patient's chest that explains the chest pain) or results of specialized tests (eg, troponin levels) that are not present for all patients and thus cannot be included in a broadly applicable algorithm. Thus, while algorithms can substantially improve decision-making, the inherent limitations of the data they learn from means that physicians' knowledge and insights will remain an important complement to these powerful tools.

ARTICLE INFORMATION

Corresponding Author: Katherine Baicker, PhD, Harris School of Public Policy, University of Chicago, 1307 E 60th St, Chicago, IL (kbaicker@uchicago.edu).

Author Affiliations: Harris School of Public Policy, University of Chicago, Chicago, Illinois (Baicker); School of Public Health, University of California, Berkeley (Obermeyer).

Conflict of Interest Disclosures: Dr Baicker reported receiving grants from the National Institute on Aging (P01 AG005842) during the submitted work; and serving on the board of directors for Eli Lilly and Mayo Clinic outside the submitted work. Dr Obermeyer reported receiving research grants from the Chan Zuckerberg Biohub, Google, the Gordon and Betty Moore Foundation, Ken Griffin (private philanthropy), the National Institutes of Health, and Schmidt Futures; receiving speaking or consulting fees from AcademyHealth, Anthem, the Attorney General of California, Blue Cross Blue Shield Tennessee, the Health Management Academy, Independence Blue Cross, and Premier Inc; having equity interests in Dandelion Health and LookDeep Health; and receiving compensation as a staff physician at Tséhootsooí Medical Center outside the submitted work.

REFERENCES

1. Mullainathan S, Obermeyer Z. Diagnosing physician error: a machine learning approach to low-value health care. *Q J Econ*. Published online December 3, 2021. doi:10.1093/qje/qjab046

2. Khullar D, Jena AB. "Natural experiments" in health care research. *JAMA Health Forum*. 2021;2(6):e210290. doi:10.1001/jamahealthforum.2021.0290

3. Zaslavsky AM. Exploring potential causal inference through natural experiments. *JAMA Health Forum*. 2021;2(6):e210289. doi:10.1001/jamahealthforum.2021.0289

4. Abaluck J, Agha L, Kabrhel C, Raja A, Venkatesh A. The determinants of productivity in medical testing: intensity and allocation of care. *Am Econ Rev*. 2016;106(12):3730-3764. doi:10.1257/aer.20140260

5. Chan DC, Gentzkow M, Yu C. Selection with variation in diagnostic skill: evidence from radiologists. *Q J Econ*. Published online January 21, 2022. doi:10.1093/qje/qjab048

6. Chandra A, Staiger DO. Identifying sources of inefficiency in healthcare. *Q J Econ*. 2020;135(2):785-843. doi:10.1093/qje/qjz040

7. Baicker K, Mullainathan S, Schwartzstein J. Behavioral hazard in health insurance. *Q J Econ*. 2015;130(4):1623-1667. doi:10.1093/qje/qjv029

8. Kolstad JT. Information and quality when motivation is intrinsic: evidence from surgeon report cards. *Am Econ Rev*. 2013;103(7):2875-2910. doi:10.1257/aer.103.7.2875