HHS Public Access

Author manuscript

Science. Author manuscript; available in PMC 2019 June 10.

Published in final edited form as:

Science. 2019 February 22; 363(6429): 810–812. doi:10.1126/science.aaw0029.

Regulation of predictive analytics in medicine:

Algorithms must meet regulatory standards of clinical benefit

Ravi B. Parikh[1], **Ziad Obermeyer**[2], and **Amol S. Navathe**[1,3]

[1]Perelman School of Medicine and Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, PA, USA.

[2]Division of Health Policy and Management, School of Public Health, University of California at Berkeley, Berkeley, CA, USA.

[3]Corporal Michael J. Crescencz VA Medical Center, Philadelphia, PA, USA.

Artificial intelligence (AI) and increased computing power have long held the promise of improving prediction and prognostication in health care (1). Now, use of predictive analytics and AI in medicine, though with fits and starts, is transitioning from hype to reality: Several commercial algorithms have received regulatory approval for broad clinical use. But the barrier for entry of new advanced algorithms has been low. To unlock the potential of advanced analytics while protecting patient safety, regulatory and professional bodies should ensure that advanced algorithms meet accepted standards of clinical benefit, just as they do for clinical therapeutics and predictive biomarkers. External validation and prospective testing of advanced algorithms are clearly needed (2), but recent regulatory clearances raise concerns over the rigor of this process. Given these concerns, we propose five standards to guide regulation of devices based on predictive analytics and AI. Although well-established research standards, such as the TRIPOD Checklist, exist for developing and validating multivariable prediction models in medicine (3), our standards provide regulatory guidance for such algorithms prior to implementation in clinical settings.

Previous generations of algorithms were largely rule-based models, often requiring manual input of usually <10 variables, to provide clinical decision support for specific situations, such as guiding imaging for pulmonary embolism, with reasonable discrimination and calibration. Over the past 5 years, modern AI-based algorithms have enabled automated real-time prediction based on almost unlimited numbers of variables, with predictive performance superior to that of traditional algorithms. Yet regulatory standards for assessing algorithms' safety and impact have not existed until recently. Furthermore, evaluations of these algorithms, which are not as readily understandable by clinicians as previous algorithms, are not held to traditional clinical trial standards. As such, there has been little prospective evidence that predictive analytics improve patient care.

Recent clearances of algorithms demonstrate the limitations of current regulatory standards. In a sentinel event, the U.S. Food and Drug Administration (FDA) in early 2018 provided

ravi.parikh@uphs.upenn.edu.

premarket clearance for the WAVE Clinical Platform, an early-warning system that integrates real-time vital sign data to identify hospitalized patients at risk of vital sign instability (4). WAVE was the first predictive surveillance platform to receive FDA clearance for clinical practice, and one of the first AI algorithm products using electronic health record (EHR) data to be cleared by a global regulatory body for widespread clinical use based on prospective evidence. Subsequent FDA device clearances for advanced algorithms, primarily in diagnostic imaging analysis, foreshadow increasing clinical availability of advanced analytics in clinical practice.

However, existing FDA standards do not neatly translate to advanced predictive algorithms. Unlike a drug or device, algorithms are not static products. Their inputs, often based on thousands of variables, can change with context. And their predictive performance may change over time as the algorithm is exposed to more data. Fortunately, the FDA has been through a similar process of developing regulatory frameworks for new diagnostics. As the era of genomic medicine began in the early 2000s, scientists discovered thousands of biomarkers purported to correlate with clinical disease. In response, organizations like the U.S. National Cancer Institute developed standards for incorporating biomarker studies into early-phase clinical trials (5). These standards included biological plausibility, validation of assay analytical performance (e.g., accuracy, specificity), specimen collection and storage standards, and laboratory needs. These eventually were formalized into an FDA Biomarker Qualification Program, a set of resources for validation of biomarkers to guide drug development and testing.

Our five criteria offer the beginnings of a similar framework for evaluation and regulation of predictive algorithms. Although not exhaustive, these criteria can not only improve the quality of predictive algorithms overall, but also ensure that these algorithms improve clinical outcomes when implemented in practice. The WAVE example, along with other exemplary models along the five dimensions, provide important lessons that validate the feasibility of these standards.

MEANGINGFUL ENDPOINTS

Although most evaluations of algorithms in medicine frame performance in abstract metrics like area under the curve (for accuracy), such metrics are not readily understandable by clinicians or patients and often are not clinically meaningful. Future evaluations should assess algorithm performance using established standards of clinical benefit. These would include downstream outcomes such as overall survival, but could also encompass other clinically relevant metrics like positive predictive value (yield of testing), number of misdiagnoses (sensitivity), and further diagnostic test characteristics.

The FDA approved the WAVE predictive algorithm on the basis of prospective data showing that the test could detect impending vital sign instability. In an early-phase study in 326 hospitalized patients, the predictive algorithm triggered an alert on average 6.3 hours before documented vital sign abnormalities (6). In a subsequent pre-post analysis without a control group, the alert was tied to a nurse-led, rapid-response intervention; this analytics-based

intervention led to a reduction in the average duration of vital sign instability per patient (16 min preintervention versus 7 min postintervention) (7).

Although the FDA's approval of WAVE was reasonable under existing regulatory standards, it also highlights their inadequacy. The manufacturer defined "clinical instability" as a deviation from normal vital signs, but is this a meaningful metric? And would a reduction in instability translate to improved patient outcomes like survival and length of hospitalization? It is unclear whether the endpoints of the WAVE evaluations were good surrogates for endpoints with clear clinical relevance, such as overall survival or rates of care utilization. Such meaningful outcomes should be emphasized in future studies of predictive algorithms —especially because algorithm outputs will be used to justify expensive and resource-intensive care for some patients and not others. As it does for drug approvals, the FDA should rigorously validate surrogate endpoints in prospective evaluations of advanced algorithms, to avoid bringing algorithms with questionable effectiveness to market.

Importantly, this does not mean that "process" metrics could not be a viable standard for FDA clearance: If a predictive algorithm could reduce providers' time spent synthesizing and interpreting complex EHR data, these intrinsically important measures could be useful for premarketing authorization. However, only a minority of algorithms receiving regulatory clearance report such process outcomes.

As regulators and professional bodies decide on which downstream outcomes matter for predictive algorithms, they should also keep in mind that some predictive algorithms—particularly those based on subjective clinician data, or outcomes that depend on access to health care—could systematically bias against certain groups of patients (8). Clinicians' responses to such biased outputs could perpetuate existing bias—and possibly harm patients. In addition to efficacy metrics, evaluations of algorithms should measure the impact of algorithm-driven interventions on care for groups at risk for this bias.

APPROPRIATE BENCHMARKS

Products based on predictive algorithms are almost never evaluated against a standard of care. As standards for comparison are not well-defined in the FDA's premarket clearance program, studies can be conducted without comparing to clinicians' predictions or guideline-based prediction scores. In a rare example of appropriate benchmarking, a deep learning algorithm recently received FDA clearance based on its ability to diagnose stroke on computed tomography imaging more rapidly than neuroradiologists. Such a comparative standard should be followed in other algorithm approvals to clarify the added value that complex and often expensive algorithms provide (9).

When benchmarking advanced predictive models, it is difficult to ensure that machine learning algorithms account for counterfactuals. For example, an algorithm that is trained on observational data of patients with sepsis may identify sepsis more accurately than physicians do, and thus will appear to be superior to physician decision-making. But this algorithm will not be trained on potential cases of sepsis that were prevented by a clinician's decision to give antibiotics. We can never know the performance of the algorithm in these

cases, which will not be in the training dataset. Hence, if an algorithm is not trained on all appropriate data, it is impossible to justify that an algorithm will improve on or should replace clinician judgment for a particular prediction. When benchmarking an algorithm's predictions against clinicians' best judgment, it is important to recognize and account for such counterfactuals by testing algorithms in multiple contexts, or perhaps including an "algorithm + clinician" arm in evaluations. By doing so, regulators may realize that such algorithms are not wholesale replacements for clinicians, but rather are complementary (or irrelevant) to clinician decision-making. Furthermore, although some evidence of the value of the predictive information can be gleaned from observational analyses of large medical databases, experimental data in the form of randomized controlled trials should be the gold standard to best assess an algorithm's value compared to routine clinical care

INTEROPERABLE, GENERALIZABLE

Algorithms receiving FDA 510(k) clearance are available for broad use. The WAVE platform algorithm is largely based on five vital signs—heart rate, respiration rate, oxygen saturation, temperature, and blood pressure—that are readily measured across health systems, and thus could be used by multiple diverse health systems. However, other complex machine learning algorithms, particularly those based on institution-specific EHR or imaging parameters, may not be easily translatable across other EHRs. In a different clinical setting, interoperability issues and unfamiliarity with the user interface may impede clinicians' ability to respond to a predictive algorithm output.

The FDA or other regulatory bodies could address this by clarifying the EHR inputs that are necessary to maximize predictive performance. This may require algorithm developers to provide extensive specification of variable inputs to ensure that commercial predictive algorithms achieve reliable and replicable results across institutions. Admittedly, regulators must balance transparency of predictive models with the proprietary interests and intellectual property of algorithm developers. However, a similar balance is commonly struck in approval of pharmaceutical agents by ensuring that pharmaceutical developers who adhere to transparency standards are given opportunity to reap the financial rewards of their product going to market.

Additionally, algorithms trained on specific populations, such as patients from a single institution, may not be generalizable across populations. A recently approved deep learning system to detect diabetic retinopathy was trained on a national database of samples from multiethnic populations, and thus is theoretically generalizable across populations (10). Training algorithms on data sources representing broad representative populations, if such data sources are available, is a model for future algorithms seeking approval across multiple settings.

SPECIFY INTERVENTIONS

Better prediction can improve quality of clinical care when tied to an intervention. Many traditional predictive rules recommended for clinical use by professional medical organizations, such as the Ottawa ankle rules or the Centor criteria, have improved practice

because they specifically guide further diagnostic workup for ankle fracture or strep throat, respectively. Recent FDA clearances, however, have not specified the interventions that should accompany an algorithm's output to improve patient care. For example, in the trial that led to its premarket authorization, the WAVE predictive algorithm triggered a nurse-led rapid response intervention; however, the details of this intervention are not mentioned in the premarket clearance notification (7). Although it may be beyond the FDA's legal purview to stipulate interventions that must accompany algorithm outputs, the FDA could provide guidance for interventions to consider when using a particular algorithm. There is precedent for this in approval of biomarker testing in areas like oncology. Standardized reporting of such interventions in clinical trials—and, in the future, registration of trials testing interventions based on predictive algorithms—may also guide clinicians and systems that are seeking to adopt such predictive algorithms.

AUDIT MECHANISMS

Just as drugs approved after clinical trials are often subject to postmarketing surveillance, predictive algorithms should be subject to rigorous audits after FDA clearance or approval. Because deep learning tools will account for new variables as time goes by, their predictive performance may change over time and over populations. An algorithm's systematic bias against certain groups may only emerge when deployed across large populations. Regular audits could help mitigate this by testing algorithmic predictions against synthetic or anonymized data. The FDA or other contracted entities could conduct such postmarketing audits without compromising intellectual property. The FDA Sentinel program for approved drugs and devices provides an example of how postmarketing audits of advanced algorithms could be accomplished with standardized claims and EHR data sources (11).

PROMISE AND PROTECTION

Modern predictive algorithms are only just beginning to be cleared by regulators and available for clinical use, so the impact of the current regulatory framework on patient outcomes is yet to be known. It is also unclear what impact the 21st Century Cures Act, which generally relaxes regulatory standards for low-risk digital health technology, will have on quality of predictive algorithms. The FDA's recent Digital Health Innovation Action Plan, issued in 2017, launched a precertification program to study clinical outcomes of AI-based tools and enable streamlined premarket review. Such efforts should be lauded but expanded upon based on our five criteria. Many developers may decry overregulation and standardization of a poorly understood field. Certainly, a commitment to regulate predictive analytics will come with time and monetary costs to these stakeholders. And policy-makers should be sensitive to the balance between regulation and innovation in this rapidly growing field. As with the field of predictive biomarkers, however, a more formal process of validating machine learning and AI can realize the promise of predictive analytics while protecting patients—moving from tremendous predictive power to improved patient outcomes.■

ACKNOWLEDGMENTS

This work was supported in part by a training grant to R.B.P. by the National Institutes of Health (5-T32-CA009615); a grant to Z.O. by the Office of the Director, National Institutes of Health (DP5 OD012161); a grant to Z.O. and A.S.N. by the Robert Wood Johnson Foundation; and a grant to A.S.N. by the Pennsylvania Department of Health, which specifically disclaims responsibility for any analyses, interpretations, or conclusions.

REFERENCES AND NOTES

1. Obermeyer Z, Emanuel EJ, Engl N. J. Med 375, 1216(2016).
2. Yu K-H, Kohane IS, BMJ Qual. Saf (2018). 10.1136/bmjqs-2018-008551
3. Collins GS, Reitsma JB, Altman DG, Moons KGM, Ann. Intern. Med 162, 55(2015). [PubMed: 25560714]
4. 510(k) Premarket Notification (2018); www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K171056.
5. Dancey JE et al.; Biomarkers Task Force of the NCI Investigational Drug Steering Committee, Clin. Cancer Res 16, 1745(2010). [PubMed: 20215558]
6. Hravnak M et al., Arch. Intern. Med 168, 1300(2008). [PubMed: 18574087]
7. Hravnak M et al., Crit. Care Med 39, 65(2011). [PubMed: 20935559]
8. Mullainathan S, Obermeyer Z, Am. Econ. Rev 107, 476(2017). [PubMed: 28781376]
9. Office of the Commissioner, Press Announcements, FDA permits marketing of clinical decision support software for alerting providers of a potential stroke in patients; www.fda.gov/newsevents/newsroom/pressannouncements/ucm596575.htm.
10. Ting DSW et al., JAMA 318, 2211(2017). [PubMed: 29234807]
11. Food and Drug Administration, Sentinel Initiative; www.sentinelinitiative.org/.