Annals of Internal Medicine

Machine Learning and Evidence-Based Medicine

Ian A. Scott, MBBS, MHA, MEd

Machine learning (ML), which converts complex data into algorithms, challenges the traditional epidemiologic approach of evidence-based medicine (EBM). Here I outline the differences, strengths, and limitations of these 2 approaches and suggest areas of reconciliation.

A Historical Context

Beginning in the 1970s, scientists extolled the virtues of EBM's hypothesis-driven, protocolized experiments involving well-defined populations and preselected exposure and outcome variables. Inferences were made using traditional biostatistics. In the early 1990s, ML emerged, whereby advanced computing programs (machines) processed huge data sets (big data) from many sources and discerned patterns among multiple unselected variables. Such patterns were undiscoverable using traditional biostatistics (1) and were used to iteratively refine (learn) layered mathematical models (algorithms). The Table lists key differences between EBM and ML.

Promise of ML Over EBM

Machine learning promises to assist clinicians in integrating ever-increasing loads of medical knowledge and patient data into routine care. Data-driven ML aims to identify similarities and differences in patient phenotypes and genomes, standardize diagnostic approaches, improve existing therapies, identify new drug targets, optimize prediction rules, avert clinical errors due to human cognitive bias and fatigue, and deliver precision medicine. Evidence-based medicine shares these goals, but ML aims to achieve them more quickly. Because it uses data sets that are already available, ML has fewer constraints related to logistics, ethics, study design, and sample size than EBM (in particular, randomized controlled trials).

Recent studies show that ML algorithms can assist clinicians in diagnosis, risk prediction, and assessment of disease severity across many medical and surgical applications (1, 2). Algorithms can match or exceed experienced clinicians in correctly diagnosing diabetic retinopathy on fundal photographs, skin cancers on dermatoscopic images, and lymph node metastases on histologic examination of biopsy specimens. Using radiologic imaging data, ML affords greater anatomical precision in administering radiotherapy for cancer and undertaking particular types of neurosurgery. Algorithms can also use clinical and laboratory data to predict surgical site infections more accurately than conventional regression models. In hospital practice, applying ML to operational data can improve efficiencies in patient triaging and appointment scheduling. Machine learning can use pharmacologic data to calculate more appropriate dosing regimens than current algorithms. In all of these cases, ML outperforms EBM and may have wider reach. Algorithms can operate at the point of care using software embedded in investigational devices, electronic health records, or mobile device applications.

Limitations of ML Versus EBM

Unlike EBM, ML algorithms often rely on routinely collected data that can be incomplete, inaccurate, subject to systematic bias, poorly described, or inaccessible, and this can lead to erroneous predictions. Seeking greater precision and external validity by using more rather than better data is problematic. Of note, ML cannot account for the frequent disagreement among clinicians about clinical features, diagnostic findings, and outcome assessments. Diverse data stored in different repositories require automated abstraction and resource-intense manual curation by experts, and unstructured notes in electronic health records are inaccessible to algorithms without layers of preprocessing (3). Natural language processing systems must expand beyond simple word recognition to incorporate semantics and syntax into their dictionaries and analyses. They must also overcome the data loss and distortion inherent in converting history and examination into a few textual notes.

Algorithms cannot recognize whether patterns or associations found in the absence of an underpinning theoretical construct are true, spurious, or affected by bias. Unlike EBM, ML has no system for rating risk of bias or quality of evidence. It cannot distinguish unwarranted from warranted practice variation and is often confounded by temporal variations in variables or sequences of clinical decisions. It may perpetuate previous errors in clinician decision making if derived from coded data reflecting prior decisions. Omission of contextual data (such as local admission policies, patient socioeconomic status, and physician preferences) may yield technically valid but misleading models (4). In contrast, EBM incorporates situational awareness and shared decision making, allowing clinicians to tailor care to context, foster relationships, and communicate findings in ways that minimize misinterpretation.

Unlike EBM, ML has limited explanatory power: Algorithms may identify many correlations between thousands of variables, but these do not prove causation. For example, intensive care admissions and inotropic infusions are highly correlated with in-hospital mortality, but stopping either or both will not prevent deaths. In EBM, randomized controlled trials demonstrate therapeutic efficacy; they also prove that ML-derived risk scores improve targeting of therapies to high-risk patients and yield better outcomes.

This article was published at Annals.org on 1 May 2018.

Table. Comparing Evidence-Based Medicine With Machine Learning

Evidence-Based Medicine	Machine Learning
Hypothesis-driven experimentation based on well-defined, sequential protocols	Data-driven discovery that uses no protocols and operates in parallel or concurrently
Examines relationships between a limited number of prespecified variables of low diversity (dimensionality)	Examines relationships between many variables that are not prespecified and have high diversity (dimensionality)
Uses structured data of lower volumes (megabytes or gigabytes), fewer participants (hundreds to thousands), and a smaller range of sources (controlled trials or prospective cohort studies)	Uses data, often unstructured, of higher volumes (terabytes or petabytes), more participants (thousands to hundreds of thousands), and a larger range of sources (EHRs, administrative data sets, wearable sensors, genomic and proteomic databanks, and social media)
Analytic methods based on theory, with declared or confirmed assumptions around data completeness, accuracy, classification, and independence	Algorithms are agnostic and data-driven, with few assumptions around data completeness, accuracy, classification, and independence
Relies on comparisons between groups to infer causation	Relies on correlations between variables within data sets to infer causation
Uses an evidence hierarchy that reflects risk of bias of specific study designs	Uses no hierarchy to assess risk of bias of different algorithms
Confidence in evidence increases with results consistently replicated in multiple studies	Confidence in algorithms developed in training sets increases with results consistently replicated in multiple testing sets

EHR = electronic health record.

Inability to see inside the "black box" of ML and understand how it arrives at results, compared with the transparency of EBM, worries clinicians (5). Even simple algorithms may fail, such as the computerized electrocardiographic readouts that misinterpreted arrhythmias (6). Opaque algorithms applied to insurance risk, employability, and other forms of social profiling have generated false and harmful predictions (7). Clinical care could incur the same problem, especially if clinicians lose skills by becoming overly reliant on automated algorithms (8). In an ML world, links between implicit and explicit knowledge that allow clinicians to imagine better ways of doing things may be lost among algorithms that simply improve the efficiency of what clinicians already do.

Reconciling EBM With ML

Despite their differences, EBM and ML can assist one another. Algorithms can facilitate more precise estimates of individual risk, with implications for choice between diagnostic tests or therapies that can then be compared in prospective, adaptive, randomized controlled trials. Regression models shown to have superior performance in ML could be applied to clinical studies that use traditional biostatistics. Mendelian randomization and statistical analyses based on directed acyclic graphs and different matching techniques may help validate causal inferences based on ML associations (9). Clinical trials, enamored of EBM, can compare ML-based interventions with usual care to assess their feasibility and validity in routine care. Hybrid algorithms are emerging that incorporate both methods and perform better than those based on 1 method alone (10). For ML to achieve "prime time" clinical application, the field needs to develop common nomenclatures, evaluation and reporting standards, comparative analyses of different algorithms, and training programs for clinicians; EBM has already traversed this path and can assist ML in doing the same.

From University of Queensland, Brisbane, Queensland, Australia (I.A.S.).

Acknowledgment: The author thanks Prof. Paul Glasziou, Director, Centre for Research in Evidence-Based Practice, Bond University, Gold Coast, Australia, and Prof. Adam Elshaug, Co-Director, Menzies Centre for Health Policy, University of Sydney, Sydney, Australia, for helpful comments on previous drafts of the manuscript.

Disclosures: The author has disclosed no conflicts of interest. The form can be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M18-0115.

Requests for Single Reprints: Ian A. Scott, MBBS, MHA, MEd, University of Queensland, Translational Research Institute, 37 Kent Street, Brisbane, Queensland 4102, Australia; e-mail, ian.scott@health.qld.gov.au.

Author contributions are available at Annals.org.

Ann Intern Med. doi:10.7326/M18-0115

References

1. **Alanazi HO, Abdullah AH, Qureshi KN.** A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care. J Med Syst. 2017;41:69. [PMID: 28285459]
2. **Hashimoto DA, Rosman G, Rus D, Meireles OR.** Artificial intelligence in surgery: promises and perils. Ann Surg. 2018. [PMID: 29389679]
3. **Bruland P, McGilchrist M, Zapletal E, Acosta D, Proeve J, Askin S, et al.** Common data elements for secondary use of electronic health record data for clinical trial execution and serious adverse event reporting. BMC Med Res Methodol. 2016;16:159. [PMID: 27875988]
4. **Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, et al.** Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. JAMA Cardiol. 2017;2:204-9. [PMID: 27784047] doi:10.1001/jamacardio.2016.3956
5. **Deo RC.** Machine learning in medicine. Circulation. 2015;132:1920-30. [PMID: 26572668] doi:10.1161/CIRCULATIONAHA.115.001593
6. **Hwan Bae M, Hoon Lee J, Heon Yang D, Sik Park H, Cho Y, Chull Chae S, et al.** Erroneous computer electrocardiogram interpretation

of atrial fibrillation and its clinical consequences. Clin Cardiol. 2012; 35:348-53. [PMID: 22644921] doi:10.1002/clc.22000

7. **Derman E.** Models.Behaving.Badly. Why Confusing Illusion With Reality Can Lead to Disaster, on Wall Street and in Life. New York: Free Pr; 2012.

8. **Cabitza F, Rasoini R, Gensini GF.** Unintended consequences of machine learning in medicine. JAMA. 2017;318:517-8. [PMID: 28727867] doi:10.1001/jama.2017.7797

9. **Linden A, Yarnold PR.** Combining machine learning and matching techniques to improve causal inference in program evaluation. J Eval Clin Pract. 2016;22:864-70. [PMID: 27353301] doi:10.1111/jep .12592

10. **Karim ME, Pang M, Platt RW.** Can we train machine learning methods to outperform the high-dimensional propensity score algorithm? Epidemiology. 2018;29:191-8. [PMID: 29166301] doi:10 .1097/EDE.0000000000000787

Author Contributions: Conception and design: I.A. Scott.
Analysis and interpretation of the data: I.A. Scott.
Drafting of the article: I.A. Scott.
Critical revision of the article for important intellectual content: I.A. Scott.
Final approval of the article: I.A. Scott.
Statistical expertise: I.A. Scott.
Administrative, technical, or logistic support: I.A. Scott.
Collection and assembly of data: I.A. Scott.