REVIEW ARTICLE

AI IN MEDICINE

Jeffrey M. Drazen, M.D., *Editor*, Isaac S. Kohane, M.D., Ph.D., *Guest Editor*, and Tze-Yun Leong, Ph.D., *Guest Editor*

Where Medical Statistics Meets Artificial Intelligence

David J. Hunter, M.B., B.S., and Christopher Holmes, Ph.D.

From the Nuffield Department of Population Health (D.J.H.) and the Department of Statistics and Nuffield Department of Medicine (C.H.), University of Oxford, Oxford, and the Alan Turing Institute, London (C.H.) — both in the United Kingdom. Dr. Holmes can be contacted at chris.holmes@stats.ox.ac.uk or at the Department of Statistics, University of Oxford, 24-29 St. Giles', Oxford OX1 3LB, United Kingdom.

N Engl J Med 2023;389:1211-9.
DOI: 10.1056/NEJMra2212850

STATISTICS EMERGED AS A DISTINCT DISCIPLINE AROUND THE BEGINNING of the 20th century. During this time, fundamental concepts were developed, including the use of randomization in clinical trials, hypothesis testing, likelihood-based inference, P values, and Bayesian analysis and decision theory.[1,2] Statistics rapidly became an essential element of the applied sciences, so much so that in 2000, the editors of the *Journal* cited "Application of Statistics to Medicine" as one of the 11 most important developments in medical science over the previous 1000 years.[3] Statistics concerns reasoning with incomplete information and the rigorous interpretation and communication of scientific findings from data. Statistics includes determination of the optimal design of experiments and accurate quantification of uncertainty regarding conclusions and inferential statements from data analysis, expressed through the language of probability.

In the 21st century, artificial intelligence (AI) has emerged as a valuable approach in data science and a growing influence in medical research,[4-6] with an accelerating pace of innovation. This development is driven, in part, by the enormous expansion in computer power and data availability. However, the very features that make AI such a valuable additional tool for data analysis are the same ones that make it vulnerable from a statistical perspective. This paradox is particularly pertinent for medical science. Techniques that are adequate for targeted advertising to voters and consumers or that enhance weather prediction may not meet the rigorous demands of risk prediction or diagnosis in medicine.[7,8] In this review article, we discuss the statistical challenges in applying AI to biomedical data analysis and the delicate balance that researchers face in wishing to learn as much as possible from data while ensuring that data-driven conclusions are accurate, robust, and reproducible.

We begin by highlighting a distinguishing feature of AI that makes it such a powerful approach while at the same time making it statistically vulnerable. We then explore three particular challenges at the interface of statistics and AI that are of particular relevance to medical studies: population inference versus prediction, generalizability and interpretation of evidence, and stability and statistical guarantees. We focus on issues of data analysis and interpretation of findings. Space constraints preclude a discussion of the important area of AI and experimental design or a deep dive into the emerging area of generative AI and medical chatbots; however, we comment on this emerging area briefly.

FEATURE REPRESENTATION LEARNING

Traditional statistical modeling uses careful hands-on selection of measurements and data features to include in an analysis — for example, which covariates to in-

clude in a regression model — as well as any transformation or standardization of measurements. Semiautomated data-reduction techniques such as random forests and forward- or backward-selection stepwise regression have assisted statisticians in this hands-on selection for decades. Modeling assumptions and features are typically explicit, and the dimensionality of the model, as quantified by the number of parameters, is usually known. Although this approach uses expert judgment to provide high-quality manual analysis, it has two potential deficiencies. First, it cannot be scaled to very large data sets — for instance, millions of images. Second, the assumption is that the statistician either knows or is able to search for the most appropriate set of features or measurements to include in the analysis (Fig. 1A).

Arguably the most impressive and distinguishing aspect of AI is its automated ability to search and extract arbitrary, complex, task-oriented features from data — so-called feature representation learning.[9-11] Features are algorithmically engineered from data during a training phase in order to uncover data transformations that are correct for the learning task. Optimality is measured by means of an "objective function" quantifying how well the AI model is performing the task at hand. AI algorithms largely remove the need for analysts to prespecify features for prediction or manually curate transformations of variables. These attributes are particularly beneficial in large, complex data domains such as image analysis, genomics, or modeling of electronic health records. AI models can search through potentially billions of nonlinear covariate transformations to reduce a large number of variables to a smaller set of task-adapted features. Moreover, somewhat paradoxically, increasing the complexity of the AI model through additional parameters, which occurs in deep learning, only helps the AI model in its search for richer internal feature sets, provided training methods are suitably tailored.[12,13]

The result is that the trained AI models can engineer data-adaptive features that are beyond the scope of features that humans can engineer, leading to impressive task performance. The problem is that such features can be hard to interpret, are brittle in the face of changing data, and lack common sense in the use of background knowledge and qualitative checks that statisticians bring to bear in deciding on a feature set to use in a model. AI models are often unable to trace the evidence line from data to features, making auditability and verification challenging. Thus, greater checks and balances are needed to ensure the validity and generalizability of AI-enabled scientific findings (Fig. 1B).[14,15]

The checking of AI-supported findings is particularly important in the emerging field of generative AI through self-supervised learning, such as large language models and medical science chatbots that may be used, among many applications, for medical note taking in electronic health records.[16] Self-supervised learning by these foundation models involves vast quantities of undocumented training data and the use of broad objective functions to train the models with trillions of parameters (at the time of this writing). This is in contrast to the "supervised" learning with AI prediction models, such as deep learning classifiers, in which the training data are known and labeled according to the clinical outcome, and the training objective is clear and targeted to the particular prediction task at hand. Given the opaqueness of generative AI foundation models, additional caution is needed for their use in health applications.

PREDICTION VERSUS POPULATION INFERENCE

AI is especially well suited to, and largely designed for, large-scale prediction tasks.[17] This is true, in part, because with such tasks, the training objective for the model is clear, and the evaluation metric in terms of predictive accuracy is usually well characterized. Adaptive models and algorithms can capitalize on large quantities of annotated data to discover patterns in covariates that associate with outcomes of interest. A good example is predicting the risk of disease.[18] However, the ultimate goal of most medical studies is not explicitly to predict risk but rather to gain an understanding of some biologic mechanism or cause of disease in the wider population or to assist in the development of new therapies.[19,20]

There is an evidence gap between a good predictive model that operates at the individual level and the ability to make inferential statements about the population.[21] Statistics is mainly concerned with population inference tasks and the generalizability of evidence obtained from one

study to an understanding of a scientific hypothesis in the wider population. Prediction is an important yet simpler task, whereas scientific inference often has a greater influence on mechanistic understanding. As Hippocrates observed, "It is more important to know what sort of person has a disease than to know what sort of disease a person has."

An example comes from the recent coronavirus disease 2019 (Covid-19) pandemic. Various prediction tools for determining whether a person has severe acute respiratory syndrome coronavirus 2 infection have been reported,[22] but moving from individual prediction to inference regarding the population prevalence and an understanding of at-risk subgroups in the population is much more challenging.[23]

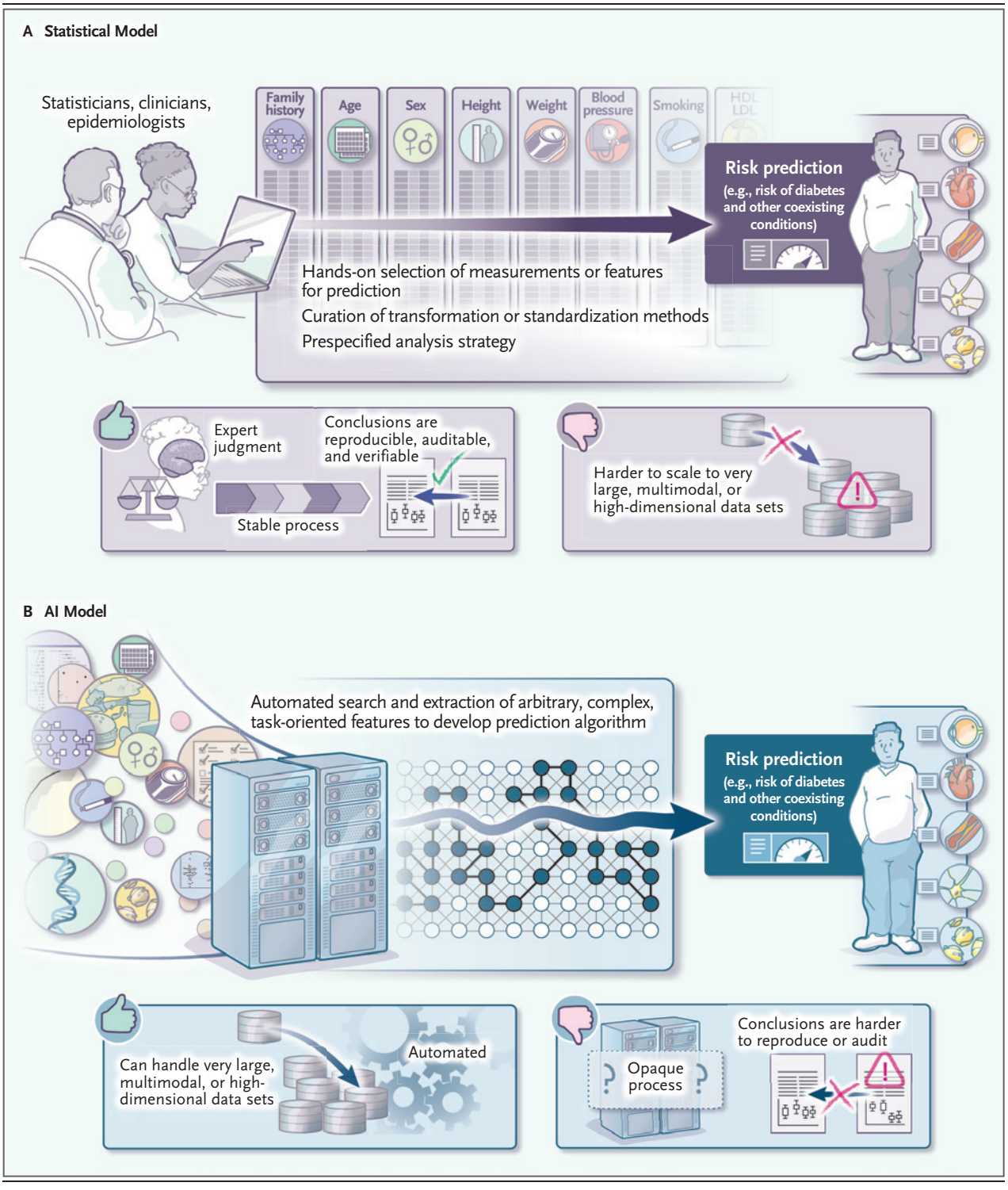
An additional challenge in the use of predictive tools is that there are many ways to measure and report predictive accuracy — for instance, with the use of measures such as the area under the receiver-operating-characteristic curve, precision and recall, mean squared error, positive predictive value, misclassification rate, net reclassification index, and log probability score. Choosing a measure that is appropriate for the context is vitally important, since accuracy in one of these measures may not translate to accuracy in another and may not relate to a clinically meaningful measure of performance or safety.[24,25] In contrast, inferential targets and estimands for population statistics tend to be less ambiguous, and the uncertainty is more clearly characterized through the use of P values, confidence intervals, and credible intervals. That said, robust, accurate, AI prediction models indicate the existence of repeatable signals and stable associations in the data that warrant further investigation.[26] Bayesian procedures have an inherent link between prediction and inference through the use of joint probabilistic modeling.[27-29]

An interesting area where AI prediction methods and statistical inference meet is causal machine learning that pays particular attention to inferential quantities.[30-32] Adoption of structural causal modeling or potential outcomes frameworks, with tools such as directed acyclic graphs, uses domain knowledge to reduce the probability that an AI model will make data-driven mistakes such as misspecifying the temporal relationship between exposure and outcome, conditioning on a variable that is caused by both exposure and disease (a "collider"), or highlighting a spurious association — for example, a batch effect in a biomarker study.[33] Causal inference methods may also be applied to AI for the interpretation of radiologic or pathological images[34] and for clinical decision making and diagnosis,[35] and they may facilitate the handling of high-dimensional confounders.[36] Although AI methods may automate and assist in applying causal inference methods to biomedical data, human judgment is likely to be necessary for the foreseeable future, if only because different AI algorithms may present us with different conclusions. Moreover, in order to avoid potential bias arising from ascertainment, mediation, and confounding, causal analysis from observational data requires assumptions that lie outside that which is learnable from the data.

GENERALIZABILITY AND INTERPRETATION

One challenge in interpreting AI results is that algorithms for internal feature representation are designed to automatically adapt their complexity to the task at hand, with nearly infinite flexibility in some approaches. This flexibility is a great strength but also requires care to avoid overfitting to data. The use of regularization and controlled stochastic optimization of model parameters during training can help prevent overfitting but also means that AI algorithms have poorly defined notions of statistical degrees of freedom and the number of free parameters. Thus, traditional statistical guarantees against overoptimism cannot be used, and techniques such as cross-validation and held-out samples to mimic true out-of-sample performance must be substituted, with the trade-off that the amount of data available for discovery is reduced. With these factors taken together, the risk is overinterpretation of the generalizability and reproducibility of results.

Practices that medical scientists should pay careful attention to in planning AI-enabled studies include releasing all code and providing clear statements on model fitting and held-out data used for reporting of accuracy so as to facilitate external assessment of the reproducibility of findings.[15] A recent report by McKinney et al. on the use of AI for predicting breast cancer on the basis of mammograms[37] prompted a call by Haibe-Kains et al. for greater transparency: "In their



study, McKinney et al. showed the high potential of AI for breast cancer screening. However, the lack of details of the methods and algorithm code undermines its scientific value."[38] The use of traditional statistical prediction methods alongside interpretable AI methods can contribute to an understanding of the prediction signal and can mitigate nonsensical associations. The

Figure 1 (facing page). Characteristics of Statistical and Artificial Intelligence (AI) Models.

As shown in Panel A, statisticians, in conjunction with clinicians, can use expert judgment to design studies and analyze the resulting data. To avoid "data dredging," the analysis is often prespecified in a statistical analysis plan, which may include such details as a listing of primary and secondary hypotheses and specification of variables that will be controlled for, how the variables will be categorized, which statistical methods will be used, how they will provide protection against type 1 error, and even how the tables will be presented. Additional or post hoc analyses are considered to be exploratory. A second statistician or statisticians starting with the same data and statistical analysis plan should produce almost identical results. These principles are challenged by high-dimensional data (i.e., data with many variables from multiple sources), for which there may be numerous alternative approaches to reducing the data to a smaller number of variables, with many options for analyses, and the statistician may "drown" in the data. As shown in Panel B, an AI algorithm can sift through vast amounts of data, but the way in which findings are derived from the data may be opaque, and it may be impossible for an analyst starting with the same data to succinctly describe and reproduce the analysis and results. Of most concern is the possibility of overfitting and of false positive results leading to findings that are not reproducible. Biases in the data that a human may understand may not be known to an AI algorithm. Internal reproducibility should be assessed by using methods such as partitioning the data into discovery and test sets. The generalizability to other data sets may be limited by idiosyncrasies in the first data set that are not shared with apparently similar data sets.

clear reporting of results and availability of code add to the potential for external replication and refinement by other groups but may be limited by a tendency to seek intellectual property rights for commercial AI products.

AI approaches may be useful in winnowing down a data set with a very large number of features, such as "-omic" data sets (e.g., metabolomic, proteomic, or genomic data), into a smaller number of features that can then be tested with the use of conventional statistical methods. Popular AI methods such as random forests, XGBoost, and Bayesian additive regression trees[39-41] all provide "feature relevance" ranking of covariates, and statistical methods such as the least absolute shrinkage and selection operator[42] use explicit variable selection as part of the model fitting. Although many AI procedures may not effectively distinguish between highly correlated variables, standard regression techniques with a smaller number of AI-selected features may do so. Feature reduction also helps the human analyst examine the data and apply additional constraints on an analysis that are based on previous subject knowledge. For example, feature A is often confounded by feature X, or a latency period of several years between exposure to feature A and the disease outcome means that no relationship is expected in early follow-up. Some similarities and differences between AI and conventional statistics are summarized in Table 1.

AI approaches challenge some recent trends in conventional statistical analysis of clinical and epidemiologic studies. Randomized trials of investigational drugs have been held to a high standard of rigor, and concerns about overinterpretation of the results of secondary end-point and subgroup analyses have led to an even stronger focus on prespecified description of primary hypotheses and control of the familywise error rate in order to limit false positive results. Protocols now often specify the precise estimands and methods of analysis that will be used to obtain P values for inference and may include the covariates to be controlled for and the dummy tables that will be filled in once data are complete. Analyses in observational studies are usually less rigorously prespecified, although a statistical analysis plan established before the start of data analysis is increasingly expected as supplementary material in published reports.[43]

AI approaches, in contrast, often seek patterns in the data that are not prespecified, which is one of the strengths of such approaches (as discussed above), and thus the potential for false positive results is increased unless rigorous procedures to assess the reproducibility of findings are incorporated. New reporting guidelines and recommendations for AI in medical science have been established to ensure greater trust and generalizability of conclusions.[44-48] Moreover, highly adaptive AI algorithms inherit all the biases and unrepresentativeness that might be present in the training data, and in using black-box AI prediction tools, it can be difficult to judge whether predictive signals arise as a result of confounding from hidden biases in the data.[49-52] Methods from the field of explainable AI (XAI) can help counter opaque feature representation learning,[53] but for applications in which safety is a critical issue, the black-box nature of AI models warrants careful consideration and justification.[54]

Table 1. Similarities and Differences between Artificial Intelligence and Conventional Statistics.

Feature	Artificial Intelligence Methods	Conventional Statistical Methods
Prior hypotheses	Agnostic or very general	Specific; often categorized as primary, secondary, and exploratory
Techniques (examples)	Random forests, neural networks, XGBoost	Parametric and nonparametric comparisons between groups; regression and survival models with linear predictors
Stability (end-to-end)	Analyses are more prone to instability and variability as a result of application domains (e.g., multimodal data integration) and user choices in algorithm specification (e.g., architecture in deep learning)	Stable analyses that follow prespecification of a statistical analysis plan with minimal available user-defined choices in model specification
Applications	Analysis of images, outputs from monitors, massive data sets (e.g., electronic health records, natural language processing)	Data with a smaller number of predictors, tabular data, randomized trials
Purpose	Pattern discovery; automatic feature representation; feature reduction to a smaller, more manageable set; prediction models	Statistical inference and testing of specific factors for departure from a null hypothesis, control of confounding and ascertainment bias, quantification of uncertainty
Reproducibility	Often internal (i.e., performed with original data set); cross-validation or split samples	Ideally external (i.e., performed with "new" data); formal tests of significance against null hypotheses
Barriers	Increasingly, use of proprietary algorithms not available to other researchers; lack of clarity in reporting	Slow progress in sharing of primary data to allow others to check or extend results
Interpretability	Often black-box; automatic algorithmic feature engineering introduces opaqueness	Explicit features, clear number of free parameters and degrees of freedom
Equity	Data-driven feature learning susceptible to biases present in data, compounding health inequities	Less flexible, more explicit (interpretable) models, which are more easily checked for equity if relevant data are available

Obermeyer and colleagues[55] describe an AI-informed algorithm that was applied to a population of 200 million persons in the United States each year in order to identify patients who were at highest risk for incurring substantial health care costs and to refer them to "high-risk care management programs." Their analysis suggested that the algorithm unintentionally discriminated against Black patients. The reason appears to be that at every level of health care expenditure and age, Black patients have more coexisting conditions than White patients do but may access health care less frequently. Thus, the algorithm with an objective function that set out to predict health care utilization on the basis of previous costs did not recognize race-related disparities in health care needs. In the future, AI algorithms may be sufficiently sophisticated to avoid this sort of discrimination, but this example illustrates both the need for human experts in clinical practice and health care policy to explore the consequences of AI applications in these domains and the need to carefully specify objective functions for training and evaluation.

STABILITY AND STATISTICAL GUARANTEES

Medical science is an iterative process of observation and hypothesis refinement with cycles of experimentation, analysis, and conjecture, leading to further experiments and ultimately toward a level of evidence that refutes existing theories and supports new therapies, lifestyle recommendations, or both. Analytic methods, including traditional statistical and AI algorithms, are used to enhance the efficiency of this scientific cycle. The context and consequences of decisions made on the basis of evidence reported in medical studies carry with them important implications for the health of patients.

To a large extent, the concern about preventing false positive results in conventional medical statistics centers on the potential clinical consequences of such results. For example, patients may be harmed by the licensing of a drug that has no benefit and may have adverse effects. In genetic analyses, falsely concluding that a chromosomal segment or a genetic variant is associated with a

disease can lead to much wasted effort attempting to understand the causal association. For this reason, the field has insisted on high LOD (logarithm of the odds) scores for linkage and very small P values for an association in genomewide studies as evidence that the association is a priori likely to represent a true positive result. In contrast, if data are being analyzed to decide whether one should be shown a particular advertisement on a browser, even a small improvement in random assignment is an improvement, and a mistake imposes a financial penalty only on the advertiser.

This difference between statistical analysis in medicine and AI analysis has consequences for the potential of AI to affect medical science, since most AI methods are designed outside of medicine and have evolved to improve performance in nonmedical domains (e.g., image classification of house numbers for mapping software[56]). In medical science, the stakes are higher, either because the conclusions may be used in the clinic or, at a minimum, false positive results will drain scientific resources and distract scientists. Trust in the robustness and stability of analyses and reporting is vital in order for the medical science community to proceed efficiently and safely. Stability refers to the end-to-end variability in the analysis, by persons skilled in the art of analysis, from project conception to end-user reporting or deployment. AI-enabled studies are increasing in complexity with the integration of multiple data techniques and data fusion. Thus, assessment of the end-to-end stability of the analysis that includes data engineering, as well as model choice, becomes vital.[57,58]

Methods that provide statistical guarantees for AI findings, such as in subgroup analysis in randomized trials[59] or observational studies,[60] can help. In the emerging area of machine learning operations, which combines machine learning, software development, and information technology operations, particular attention is paid to the importance of data engineering in the AI development cycle[61] and the problem of "garbage in, garbage out," which can affect automated machine learning in the absence of careful human intervention.

There are many examples of data analysis in medical science in which we undertake an "agnostic" analysis because a specific hypothesis does not exist, or if it does, it is global (e.g., some genetic variants among the very large number being tested are associated with the disease of interest). This obviously leads to a substantial multiplicity problem. Multiplicity can be controlled by using standard approaches such as the Bonferroni correction or explicitly using a Bayesian prior specification on hypotheses, but new AI approaches to graphical procedures for controlling multiplicity are being developed.[62] Another standard approach is to validate findings in an independent data set on the basis of whether the AI predictions are reproduced. Where such independent validation is not possible, we must resort to mimicking this approach by using in-sample partitioning. Dividing the data into two sets, one for discovery and one for validation, can provide statistical guarantees on discovery findings.[63] More generally, multiple splits with the use of cross-validation can estimate future predictive risk,[64] although statistical uncertainty in the predictive risk estimate is harder to assess. Emerging techniques in conformal inference look promising for quantifying uncertainty in prediction settings.[65]

STATISTICAL SENSE AND THE ART OF STATISTICS

Much of the art of applied statistics and the skills of a trained statistician or epidemiologist involve factors that lie outside the data and, hence, cannot be captured by data-driven AI algorithms alone. These factors include careful design of experiments, an understanding of the research question and study objectives, and tailoring of models to the research question in the context of an existing knowledge base, with ascertainment and selection bias accounted for and a healthy suspicion of results that look too good to be true, followed by careful model checking. Bringing these skills to bear on AI-enabled studies through "human-in-the-loop" development (in which AI supports and assists expert human judgment) will enhance the effect and uptake of AI methods and highlight methodologic and theoretical gaps to be addressed for the benefit of medical science. AI has much to bring to medical science. Statisticians should embrace AI, and in response, the field of AI will benefit from increased statistical thinking.

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

REFERENCES

1. Gorroochurn P. Classic topics on the history of modern mathematical statistics: from Laplace to more recent times. Medford, MA: Wiley, 2016.
2. Wasserman L. All of statistics. New York: Springer, 2013.
3. Looking back on the millennium in medicine. N Engl J Med 2000;342:42-9.
4. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380:1347-58.
5. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44-56.
6. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28:31-8.
7. Parikh RB, Obermeyer Z, Navathe AS. Regulation of predictive analytics in medicine. Science 2019;363:810-2.
8. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019;25:30-6.
9. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 2013;35:1798-828.
10. Schölkopf B, Locatello F, Bauer S, et al. Toward causal representation learning. Proc IEEE 2021;109:612-34.
11. Choi E, Bahadori MT, Searles E, et al. Multi-layer representation learning for medical concepts. In: Proceedings and abstracts of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13–17, 2016. San Francisco: Special Interest Group on Knowledge Discovery and Data Mining, 2016.
12. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
13. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA: MIT press, 2016.
14. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17:195.
15. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 2020;368:l6927.
16. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388:1233-9.
17. Agrawal A, Gans J, Goldfarb A. Prediction machines: the simple economics of artificial intelligence. Cambridge, MA: Harvard Business Press, 2018.
18. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J 2017;38:1805-14.
19. Bzdok D, Engemann D, Thirion B. Inference and prediction diverge in biomedicine. Patterns (N Y) 2020;1:100119.
20. Cox DR. Statistical modeling: the two cultures. Stat Sci 2001;16:216-8.
21. Bzdok D, Ioannidis JPA. Exploration, inference, and prediction in neuroscience and biomedicine. Trends Neurosci 2019;42:251-62.
22. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020;369:m1328.
23. Rossman H, Segal E. Nowcasting the spread of SARS-CoV-2. Nat Microbiol 2022;7:16-7.
24. Oren O, Gersh BJ, Bhatt DL. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. Lancet Digit Health 2020;2(9):e486-e488.
25. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet 2019;393:1577-9.
26. Sammut S-J, Crispin-Ortuzar M, Chin S-F, et al. Multi-omic machine learning predictor of breast cancer therapy response. Nature 2022;601:623-9.
27. Rubin DB. Bayesian inference for causal effects: the role of randomization. Ann Stat 1978;6:34-58.
28. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. New York: Taylor and Francis, 2014.
29. Fong E, Holmes C, Walker SG. Martingale posterior distributions. November 22, 2021 (https://doi.org/10.48550/arXiv.2103.15671). preprint.
30. Peters J, Janzing D, Schölkopf B. Elements of causal inference: foundations and learning algorithms. Cambridge, MA: MIT Press, 2017.
31. Van der Laan MJ, Rose S. Targeted learning: causal inference for observational and experimental data. New York: Springer, 2011.
32. Blakely T, Lynch J, Simons K, Bentley R, Rose S. Reflection on modern methods: when worlds collide-prediction, machine learning and causal inference. Int J Epidemiol 2021;49:2058-64.
33. Sanchez P, Voisey JP, Xia T, Watson HI, O'Neil AQ, Tsaftaris SA. Causal machine learning for healthcare and precision medicine. R Soc Open Sci 2022;9:220638.
34. Castro DC, Walker I, Glocker B. Causality matters in medical imaging. Nat Commun 2020;11:3673.
35. Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. Nat Commun 2020;11:3923.
36. Clivio O, Falck F, Lehmann B, Deligiannidis G, Holmes C. Neural score matching for high-dimensional causal inference. Presented at the 25th International Conference on Artificial Intelligence and Statistics, virtual, March 28–30, 2022. poster (https://virtual.aistats.org/virtual/2022/poster/3447).
37. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577:89-94.
38. Haibe-Kains B, Adam GA, Hosny A, et al. Transparency and reproducibility in artificial intelligence. Nature 2020;586(7829):E14-E16.
39. Breiman L. Random forests. Mach Learn 2021;45:5-32.
40. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. Ann Appl Stat 2010;4:266-98.
41. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings and abstracts of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13–17, 2016. San Francisco: Special Interest Group on Knowledge Discovery and Data Mining, 2016.
42. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996;58:267-88.
43. Harrington D, D'Agostino RB Sr, Gatsonis C, et al. New guidelines for statistical reporting in the *Journal*. N Engl J Med 2019;381:285-6.
44. Sounderajah V, Ashrafian H, Aggarwal R, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI Steering Group. Nat Med 2020;26:807-8.
45. CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. Nat Med 2019;25:1467-8.
46. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 2020;26:1320-4.
47. DECIDE-AI Steering Group. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. Nat Med 2021;27:186-7.
48. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020;368:m689.
49. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021;27:2176-82.
50. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. JAMA 2019;322:2377-8.
51. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias

and fairness in machine learning. ACM Comput Surv 2021;54:1-35.
52. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. Annu Rev Biomed Data Sci 2021;4:123-44.
53. Tjoa E, Guan C. A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE Trans Neural Netw Learn Syst 2021;32:4793-813.
54. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206-15.
55. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366:447-53.
56. Deng L. The MNIST database of handwritten digit images for machine learning research. IEEE Signal Process Mag 2012;29:141-2.
57. Yu B, Kumbier K. Veridical data science. Proc Natl Acad Sci U S A 2020;117:3920-9.
58. Gelman A, Loken E. The statistical crisis in science: data-dependent analysis — a "garden of forking paths" — explains why many statistically significant comparisons don't hold up. Am Sci 2014;102:460 (https://www.americanscientist.org/article/the-statistical-crisis-in-science).
59. Watson JA, Holmes CC. Machine learning analysis plans for randomised controlled trials: detecting treatment effect heterogeneity with strict control of type I error. Trials 2020;21:156.
60. Nie X, Wager S. Quasi-oracle estimation of heterogeneous treatment effects. Biometrika 2021;108:299-319.
61. Treveil M, Omont N, Stenac C, et al. Introducing MLOps. Newton, MA: O'Reilly, 2020.
62. Zhan T, Hartford A, Kang J, Offen W. Optimizing graphical procedures for multiplicity control in a confirmatory clinical trial via deep learning. Stat Biopharm Res 2022;14:92-102.
63. Cox DR. Some problems connected with statistical inference. Ann Math Stat 1958;29:357-72.
64. Stone M. Cross-validatory choice and assessment of statistical predictions. J R Stat Soc Series B Stat Methodol 1974;36:111-47.
65. Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L. Distribution-free predictive inference for regression. J Am Stat Assoc 2018;113:1094-111.