9 Biostatistics and Artificial Intelligence

Lance A. Waller

Introduction

We live in an increasingly data-rich world. Multiple, linked devices track our location and put us into the context of the world around us. We post and comment on our activities, our photos, and our friends. Our constantly connected world provides benefits (real-time driving directions) and risks (a loss of privacy). In the world of clinical medicine and biomedical research, rapid expansion in measurement technology generates data in volumes and at speeds previously unimaginable. Detailed medical images are created, collated, and stored across a wide range of patients every day. Data, formerly only associated with scientific inquiry, now surround us and drive our daily decisions.

Not only is the nature and availability of data changing, but our approaches to managing, analyzing, and understanding data are shifting to accommodate novel data settings. The discipline of *data science* includes familiar elements of mathematics, statistics, and computing but places these ideas in new settings and scenarios. This expanded framework extends the data analytic toolbox to include *machine learning* and other artificial intelligence approaches building on newly available data, newly available computational platforms, and newly available storage capacities.

How do new data types, new computational platforms, and new analytic techniques fit into biomedical imaging research? In this chapter, we provide a brief overview of traditional and emerging data types and outline how tools, algorithms, and ideas from the fields of biostatistics and machine learning, separately and together, enhance the analytic toolbox available in cardiothoracic imaging research.

Shifting Data Paradigms and Analytic Paradigms

Many years ago, researchers viewed scientific data as expensive, particularly biomedical data from clinical settings. In this *traditional data paradigm*, each research project collected its own relevant data with a goal of gaining the most information from the smallest sample size that provided a prespecified level of statistical accuracy. In contrast, the rapid growth of biomedical measurement technology, computational processing power, and computational storage provide researchers ready access to large sets of measurements. In this *emerging data paradigm*, available data may not include *exactly* what a biomedical investigator wishes to collect, but there are large amounts of related measurements, similar to what is desired. For instance, clinical settings often now routinely store and retain large numbers of digitized images (scanned from historical film images or collected directly in a digital format), thereby providing a collection of images far greater than can be collected within the typical budgetary constraints of a single research project. From an investigator's perspective, the traditional and emerging paradigms reflect a difference in focus between "What specific data should I gather?" and "What relevant data can I find?" The rapid increase in the availability of and access to such data repositories motivates investigators to wonder: At what point does the volume of available-but-imperfect available data outweigh the focused design of investigator-based collection of data specifically tailored to address a research question?

Accompanying this changing data paradigm is a shift in the data analysis paradigm from a primary focus on *estimation* of associations between interventions and outcomes to *prediction* of outcomes given a set of observations (including interventions). Both estimation and prediction have been present in thinking about data analysis for some time [4], but as greater amounts of data become more available and accessible, we find an increasing focus on prediction. The

L. A. Waller, BS, PhD (✉)
Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA, USA
e-mail: lwaller@emory.edu

C. N. De Cecco et al. (eds.), *Artificial Intelligence in Cardiothoracic Imaging*, Contemporary Medical Imaging,
https://doi.org/10.1007/978-3-030-92087-6_9

emerging data paradigm, coupled with large-scale cloud computing, allows the development of data-hungry prediction algorithms exploring and classifying massive data sets with impressive results. Familiar examples include crowdsourced calculations of expected travel times utilizing volunteered location data from millions of mobile phones to online translators matching phrases in one language to their best match in another. Looking further, image classifiers can identify facial features or potential malignancies, and autonomous vehicles are "driven" by algorithms that rapidly classify the input from multiple sensors to predict the best course of action. Networks of finely tuned prediction algorithms can learn from large-scale collections of data and form the basis of many current artificial intelligence (AI) systems controlling everything from our household thermostats to complex communications systems.

The success of AI systems, particularly those based on variants of *machine learning* algorithms, prompts some to propose that we are headed toward "hypothesis-free science" or to boldly state that all fields of science will be branches of computer science in the near future [3]. However, while such statements typically reflect self-promotional hyperbole, it is important to assess whether artificial intelligence based on machine learning is replacing or simply is adding to the analytic options open to the biomedical investigator. In order to explore this question, we will compare brief overviews of the two data paradigms and accompanying data analytic paradigms mentioned above.

The Traditional Data Paradigm and Biostatistical Analysis

While probability calculations had been developed (in part to understand games of chance) in the eighteenth and nineteenth centuries, the twentieth century saw a detailed development of disciplined *experimental design* wherein data were purposefully collected in ways to explore predefined hypotheses taking into account observed variation in measurements. The traditional data paradigm (data are expensive, so focus collection on what is needed to best answer the question at hand) drove the development of statistical techniques focused on comparisons of intervention and nonintervention groups adjusting for (and attempting to minimize) the variances in measurements in both groups. The mathematical rules of probability defining the distributions of observations provided the ingredients for estimation of and inference about distributional parameters (e.g., means, variances, correlations) from observations generated from these distributions. Statistically independent observations provided the most information per observation, and sample size calculations allowed experimenters to define how many independent observations would be required to provide estimates within a given level of precision. Regression-based methods and models (e.g., linear regression, generalized linear models) provided analytic tools for estimating associations between covariates and outcomes. Independence assumptions were relaxed, allowing longitudinal correlations between observations from the same patient over time, correlations between different patients evaluated at the same clinic, or spatial correlations between observations taken at nearby locations (e.g., neighboring voxels in a medical image). The statistical toolbox is large and offers tools as simple as comparing mean values between two groups and as complex as identifying regions of significant difference between resting and active states in brain and heart images. Importantly, the biostatistical toolbox is not complete but rather grows to address the challenges that arise when traditional assumptions are violated.

The field of *biostatistics* focuses on building and expanding this traditional data toolbox; specifically, biostatistics involves the development, application, interpretation, and evaluation of statistical methods for application to biomedicine and public health. To illustrate, two general areas of biostatistical methodology research focus include clinical trials and epidemiology.

Clinical trials are defined by the US National Institutes of Health as:

> A research study in which one or more human subjects are prospectively assigned to one or more interventions (which may include placebo or other control) to evaluate the effects of those interventions on health-related biomedical or behavioral outcomes. (https://grants.nih.gov/policy/clinical-trials/definition.htm)

Several key aspects arise in this definition. First, clinical trials involve experimentation with human subjects, where interventions are prospectively assigned and outcomes are evaluated. Ethical and safety concerns require regulation and clearly designed protocols for participant recruitment, intervention assignment, measurement of outcomes, and evaluation of differences in outcomes by intervention groups.

In the clinical trial setting, experimental design is of utmost importance, and biostatistical issues play critical roles. Investigators must summarize past knowledge, define clear hypotheses in terms of anticipated intervention effects, evaluate measurement uncertainty, predefine analytic plans, and calculate appropriate sampling plans and sample sizes to recruit participants, assign interventions, collect data, and evaluate results. Central elements include summaries of measurements within groups, assessment of variation in these measurements, and comparisons of summaries between intervention groups. Comparisons are often assessed via statistical hypothesis testing or through the calculation of confidence interval estimates.

Statistical hypothesis testing and its summary of statistical significance, the p-value, have received critical consideration in recent years [1]. The p-value represents the chance of observing a summary measure at least as extreme as that observed in the data, if the null hypothesis of no intervention effect is true. That is, if we were to repeat the experiment under identical conditions and there was no difference between intervention groups, what proportion of the time would we see an observed difference in summaries at least as large as that we observed in our observed data? Note, the definition of the p-value is different than simply representing the probability that the null hypothesis is true (as many of us were reminded frequently in our statistics courses!). Much of the recent criticism of the use of *p*-values to determine whether experimental findings are important is based on the overreliance of a fixed (usually 0.05) threshold for statistical significance and using this threshold to define whether an investigator has a publishable finding. In today's data-rich world, blindly applying a p-value threshold to a large number of tests ("p-hacking") will invariably lead to a result that is statistically significant but may only be due to chance [7]. In addition, in the case of very large sample sizes, statistically significant differences may be of such small magnitude as to be clinically irrelevant.

In contrast to the experimental setting of clinical trials, the field of epidemiology often involves the analysis of observational (i.e., non-experimental) data wherein an investigator compares data from individuals with and without the health outcome of interest with respect to factors that may contribute to differences in risk. Such observational studies can be based on prospective data wherein cohorts of individuals are followed over time to observe which individuals develop the outcome of interest and which do not or can be based on retrospective data wherein investigators identify individuals with and without the outcome of interest and look back in time to evaluate past exposures that may relate to risk.

Relevant biostatistical issues of measurement variability, estimation of contrasts between groups, and assessments of similarity between intervention (or, more generally, exposure) groups remain key biostatistical issues in epidemiologic data analysis. In addition, the observational setting includes a reduction in the amount of control an investigator has over the settings and experiences of research subjects, and as a result, issues of bias, confounding, and interactions can occur between risk factors and influence (bias) estimated associations between risk factors and outcomes, an issue of increased importance in epidemiology compared to the more controlled setting of clinical trials.

The transition from the traditional data paradigm where investigators collect their own data to the emerging data paradigm where investigators both collect and seek out existing data relevant to the questions of interest has a clear impact on both clinical trials and epidemiology. In clinical trials, the availability of past clinical (and nonclinical) information can rapidly increase identification and enrollment of potential participants meeting eligibility criteria. In epidemiology, the rapid access of administrative data can, for example, reveal past patterns in housing type, environmental exposures, traffic volumes, and clinic location. Such examples suggest potential opportunities for emerging data analytic tools to better incorporate larger volumes of available and accessible data. Before considering how one might combine traditional biostatistical thinking and tools with modern data access, we first provide a brief overview of machine learning approaches to analyzing data in the emerging data paradigm.

The Emerging Data Paradigm and Machine Learning

In the emerging data paradigm, large volumes of data are available and can be accessed for analysis. In addition, this paradigm also encourages data sharing to allow reproducible research wherein results may be verified, adapted, and extended by other investigative teams. To support this ideal, increasingly, many fields of science aim to follow the FAIR data principles, i.e., making data Findable, Accessible, Interoperable, and Reusable [13]. Scientific journals increasingly require data associated with publications to be available as a condition of publication. While data sharing can be complicated with biomedical data, particularly clinical data that often fall under federal privacy regulations, in some cases, the development of clearly defined data use agreements can allow sharing of data at some level of aggregation. Simply put, the curation, storage, and citation of data sets are a growing and dynamic part of today's biomedical research environment, and the creation and curation of citable data sets are becoming a critical component of a researcher's scholarly output [11]. The broad and growing universe of generally accessible data motivates many new approaches to data analyses, including the broad field of machine learning.

As noted above, machine learning algorithms form the general basis of much of today's artificial intelligence applications in biomedicine and other fields. Such algorithms explore large, complex data sets seeking ways to classify information and "learn" how to classify the next sets of observed data. While machines do not really "learn" in the same way as living creatures, algorithms can be very efficient in classifying patterns within large data sets to build computational and statistical *models* linking input data (e.g., clinical records, images) into output labels and classifications (e.g., diagnostic categories of healthy, moderate disease, and severe disease).

In most machine learning applications, each data observation includes values for a large (sometimes *very* large) number of variables. The *models* mentioned above are typically mathematical functions that, roughly speaking, weight observed input variable to provide a summary score as output. This summary score can be a single value or a collection of values (e.g., a scorecard) but is typically of much smaller dimension than the original data. For example, in a cardiothoracic imaging study, the data from each patient may include a high-resolution image (or series of images), patient demographics, chart history, etc. The collection of input data can be heterogeneous and contain numerical values, voxel-specific readings from images, and text information from clinical notes.

Continuing the example, for *output*, we may be interested in identifying image-based predictors associated with the identification and diagnoses of a specific physiological anomaly, something that traditionally would be diagnosed by a human reader from diagnostic imaging in consultation with clinical visits. The function translating input to output will be a summary of all of the information in patient images and charts giving high weight to data elements related to the anomaly of interest and low (or zero) to data elements unassociated with the anomaly. These weights could be assigned by an expert or, in the case of machine learning, calculated by evaluating which input values are most strongly associated with the output diagnostic classification across a very large number of images. The goal is to define the function (set of weights) that can best classify new data inputs (say, new images) that do not have associated diagnosis information. In other words, we focus on *prediction* and seek to define the function that best predicts the outcome based on the input. This shift from the traditional analysis goal of *estimation* of associations between interventions and outcomes to *prediction* of outcomes given interventions is a key ingredient in the rise of *data science*. This idea is eloquently described in an overview of the "two cultures" of statistical modeling described in Breiman [4] and expanded in historical context by Donoho [6].

Machine learning builds on the input-to-function-to-output idea where our goal is to use an algorithm to identify the ingredients of our function that optimize our ability to provide the best output. Differing details defining optimization and what defines the "best" predictions lead to different flavors of machine learning, and we briefly define two general classes of algorithms below. For readers interested in broader overviews, Shane [9] provides a very readable and entertaining introduction to the intuition driving what machine learning does (and does not) do well, while Burkov's [5] text provides a more technical (but still very readable) summary of specific families of machine learning approaches.

To motivate our discussion, suppose we continue our example above: we have a large set of cardiac images and we want to build an algorithm to scan and identify the images indicating the presence of a particular anatomical defect. Such an algorithm, if perfect, would remove the need for human review of images. While no classification system (even an expert human reader) is perfect, even a system that could identify and classify the subset of images worth a more careful look by a human expert could save a great deal of time and effort in processing large numbers of images in clinical practice.

Typical machine learning algorithms will split the full set of images into a *training set* and a *validation set*. We use the training set to develop and refine our algorithm and then apply the algorithm to the validation set to see how well it does. *Supervised machine learning* begins with a set of images in the training set labeled by an expert (or set of experts) as containing the anomaly or not. The algorithm will then seek to identify which features in the labeled images have in common with each other in order to identify what information in the images represents signal related to the desired classification and what information represents noise. The algorithm optimizes the definitions of weights in the function linking input data to predicted outcomes based on the labeled images in the training set. In some settings, all images in the training set will be prelabeled, but, more commonly, the training data include both labeled and unlabeled images. If this is the case, the algorithm next applies the features identified as signal in the labeled images to the unlabeled images in the training set and begins to classify these as well. The algorithm picks up additional features from the unlabeled images that appear to contain the defect and examines whether these features are also present in the labeled features. After this classification exercise with the data in the training set, the algorithm will have a set of rules (the weights in the function) to classify images. Note that this set of rules has been developed and tested on the training set. Next, to assess whether the rules apply to data not included in the process of fine-tuning the algorithm, we apply the algorithm with the optimized weights to the data in the validation set to evaluate which images the algorithm identifies as containing the anomaly. Having human experts review the algorithm's diagnostic predictions for the validation data allows the user to assess how well the algorithm will behave on new data.

In contrast to supervised learning, *unsupervised learning* follows much the same process (develop the algorithm on the test data set and then apply to the validation set), but without the initial expert labels. In this case, the algorithm seeks to classify images by common features. Typically, unsupervised machine learning will not have a single goal in mind (e.g., identifying images containing evidence of the anomaly) but rather seeks to identify groups (often multiple groups) of images with common features. These groups of classified images will then be reviewed by subject matter experts to assess the diagnostic utility of the groups, perhaps leading to a subsequent round of supervised learning.

Extending these ideas further, if one includes multiple steps in machine learning (i.e., "learning" from the output of previous machine learning results), one may extend the ideas into the area of *deep learning* [2]. Multiple layers of optimization and evaluation can provide additional levels of precision in classification problems, much the same way that a biopsy and more detailed testing can provide additional specificity in diagnosis following a cancer screening test. Each step refines the previous step.

Discussion: Analytic Toolboxes with Both Types of Approaches

The brief descriptions above sweep many details under the rug but highlight a key difference in goals between the traditional biostatistical and the machine learning approaches to biomedical data analysis. At the risk of oversimplifying, we contrast the motivations of the two data analytic paradigms as follows:

- Statistics begins with describing how the data were collected using probability distributions and then uses the data observed to estimate the defining characteristics (parameters) of those distributions. The distributions define the expected observations (and their variability) under different intervention settings in order to assess differences in outcomes between intervention settings.
- Machine learning begins with large quantities of data and evaluates these data to optimize prediction of patterns within the data. In our examples, we place special emphasis on patterns relating to health outcomes and diagnoses. Importantly, while causal associations are often of interest, by default, patterns found in machine learning predictions need not be causal in nature to be useful for prediction; they are simply based on correlations. For instance, an increase in Internet searches for high school basketball scores is not directly causally related with increased risk of influenza, but both basketball searches and flu cases increase in the fall. While banning basketball may not measurably impact the transmission of influenza, including the temporal pattern of basketball searches may improve our predictions of influenza transmission each year.

Machine learning algorithms power much of current commercially available AI, and there is a large volume of research exploring their application to human healthcare and biomedical research. The first wave of such research often focuses on showing one *can* apply machine learning approaches to high-volume human health data, but interesting and important work on how and whether one *should* apply such approaches is starting to appear. In one example, Vollmer et al. [10] raise a set of critical questions specifically for assessing patient benefits of machine learning and AI research. These fall into six categories, briefly summarized here:

- Study inception (What information/hypotheses motivated the study?)
- Study (How do data related to the question of interest? How are patients involved?)
- Statistical methods (What analytic approaches drive study conclusions?)
- Reproducibility (Are data available to other researchers? If so, in what form?)
- Impact evaluation (How do study results generalize to the full population?)
- Implementation (Is the AI model cost effective for practical use in patient care?)

These and the full list of questions in Vollmer et al. [10] outline a principled discussion of how and where machine learning algorithms can benefit patient-based biomedical research and highlight different decisions in different dimensions of the research cycle.

In addition to these important questions across the patient-based research cycle, it is also important to recognize that while the traditional and emerging analytic paradigms are different [8], they need not be mutually exclusive. In fact, it is important to use tools from biostatistics to assess performance of machine learning, and it is important for biostatistics to adopt some of the concepts, capacity, and capability of machine learning and the emerging data paradigm.

The field of biostatistics can benefit from the possibilities enabled when one considers widely available data of different types, and the field of machine learning can benefit from traditional biostatistical understanding aggregate behavior of diagnostic decisions. To illustrate an example of the later, suppose the outcome of a machine learning algorithm is to identify the presence of a specific cardiac anomaly in biomedical imaging and that we have a very large number of historical images. Machine learning is well-suited to this problem, and once our algorithm is defined, we can input large number of new images and receive a binary (anomaly present/absent) classification for each. As noted above, no diagnostic tool (algorithm or expert) is perfectly accurate. Note that there are two types of correct diagnoses: an image from a patient who has the anomaly can be diagnosed with anomaly present (a *true positive* result), and an image from a patient who is anomaly-free is diagnosed as anomaly-free (a *true negative* result). In biostatistics, the term *sensitivity* denotes the probability of the algorithm diagnosing an anomaly present when the patient indeed has the anomaly, the probability of a true positive result. In biostatistics, the *specificity* denotes the probability of the algorithm diagnosing the absence of the anomaly when, in fact, the patient does not have the anomaly, the probability of a true negative result. Sensitivity (diagnosing presence when the anomaly is indeed present) and specificity (diagnosing absence when the anomaly is indeed not present) are two related but different measures of diagnostic performance from the biostatistical literature on diagnostic testing. Note that there are also two types of mistaken diagnosis: an image from a patient who does indeed have the anomaly can be diagnosed as anomaly-free (a *false-negative* result), and an image from an anomaly-free patient can be diagnosed as having the anomaly (a *false-positive* result).

These definitions contain two sets of related probabilities:

- If sensitivity is high, the probability of a false-negative result is low (these probabilities represent a proportion of the population of *individuals with the anomaly*). In individuals with the anomaly, the algorithm diagnoses one of two options, so the sensitivity and the probability of a false-negative result add to one.
- Similarly, if the specificity is high, the probability of a false positive is low (these probabilities represent proportions of the population of *individuals without the anomaly*). If individuals do not have the anomaly, the algorithm diagnoses one of two options, so the sum of specificity and the probability of a false-positive result is one.

It is important to note sensitivity and specificity are properties of the same diagnostic tests, but they represent proportions of *different groups of patients* and cannot be added or directly compared to each other.

Suppose we have a very large number of images available, and we have a machine learning algorithm with high sensitivity (say, 0.99) and high specificity (say, 0.95). The emerging data paradigm suggests we apply the algorithm to the set of images and collect the diagnoses. However, past biostatistical experience suggests that we may find some confusing results, depending on how common the anomaly is within our large number of images. Let us take a closer look.

For every 100 images from individuals with the anomaly, the sensitivity of 0.99 suggests we will receive, on average, one false-negative result. For every 100 images from individuals without the anomaly, the specificity of 0.95 suggests that we will receive, on average, five false-positive results. Across the 200 images combined, we will observe (on average) 104 positive results (99 true positives + five false-positive results). The *positive predictive value* (the proportion of anomaly diagnoses that are true will be 99/104 = 95.19%, a good value). If we maintain an equal balance of images from individuals with and without the anomaly, we will maintain this same high value of the positive predictive value. (To see this, note that the same result happens if we have 100,000 images with the anomaly and 100,000 without and the same with 100,000,000 images in each category.)

However, suppose the anomaly is rare in our set of images, say, only one image has the anomaly out of every ten images without the anomaly. In this case, we say the *prevalence* of the anomaly is 1/10. If we apply the algorithm to 100 images with the anomaly, we will again, on average, receive 99 true positive results and one false-negative result. The prevalence of 1/10 means that for our 100 images with the anomaly, we apply the algorithm to 1000 images without the anomaly. Based on the specificity of 0.95, we will, again, diagnose (on average) 5% of these images as false positives. Five percent of 1000 images corresponds to 50 false-positive diagnoses. In case where prevalence of the anomaly in our total set of images is 1/10, we observe, on average, 149 positive results (99 true positives and 50 false positives). In this case, the positive predictive value is 99/149 = 66.44%, a considerable reduction from the positive predictive value when half of our images were from patients with the anomaly.

If we continue to test 100 images that truly contain the anomaly but test more and more images without the anomaly (i.e., we further reduce the prevalence of images with the anomaly within the set of all images tested), we note that even though the *proportion* of false-positive diagnoses stays constant with a specificity of 95%, the *number* of false-positive results increases with decreasing prevalence, for instance:

- With 10,000 anomaly-free images, we expect 500 false-positive diagnoses.
- With 100,000 anomaly-free images, we expect 5000 false-positive diagnoses.
- With 1,000,000 anomaly-free images, we expect 50,000 false-positive diagnoses.

If we only have 100 images with anomalies in each case, we see we will soon have more false-positive diagnoses than true positive diagnoses.

This association between prevalence and positive predictive value is well-known in biostatistics and is often used to motivate the use of Bayes' theorem to link the concepts of sensitivity, specificity, prevalence, and positive predictive value. Despite its familiarity, the associations can be confusing, and multiple press reports relating to the performance of mass testing for COVID19 often expressed exasperation that a diagnostic test with high sensitivity and specificity could have such low positive predictive value [12]. The machine learning community rediscovered the association and refers to it a problem of *unbalanced design* [9]. However, sometimes, the situation is viewed as a signal to revise the algorithm with improved weights rather than recognizing that the problem is, in a sense, baked into diagnostic systems. Rather than adjusting a single algorithm, providing multiple layers of diagnoses and testing, similar to follow-up clinical testing following a screening diagnosis, offers more control of the positive predictive value and is, roughly speaking, the basis of the deep learning approaches mentioned above.

The diagnostic testing example illustrates that even in a relatively straightforward setting, both machine learning and biostatistics have much to learn from each other. A traditional data approach of only analyzing images collected in the current study would miss the opportunity to gain information from the full set of available images, while a machine-learning-only approach applied to all available images could result in very poor positive predictive performance of an automated diagnostic procedure potentially resulting in unnecessary treatment for patients.

Conclusion

In conclusion, we note that we are working in a transitional time. Amounts of and access to data are changing rapidly, and our analytic tools should both adapt to new settings and take advantage of lessons learned in the past. The evolution of data science in the biomedical research arena requires expertise from both analytic paradigms in order to gain advantage of today's data resources. The data analysis toolbox is not static, nor is it complete, and the savvy investigator will work closely with data scientists, statisticians, computer scientists, and mathematicians in order to gain the most information from the available data and minimize chances for erroneous conclusions.

In an interview with a *Saturday Evening Post* writer, bank robber Willie Sutton famously answered the question "Why do you rob banks?" with "Because that's where the money is" [14]. The data analytic parallel is our recommendation advocating the informed use of all tools available in data analysis. This allows us to answer the question "Why do you analyze data?" "Because that's where the information is."

References

1. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567:305–7.
2. Baraniuk R, Donoho D, Gavish M. The science of deep learning. Proc Natl Acad Sci USA. 2020;117:30029–32. https://doi.org/10.1073/pnas.2020596117.
3. Barker ED, Roberts S, Walton E. Hidden hypotheses in 'hypothesis-free' genome-wide epigenetic associations. Curr Opin Psychol. 2019;27:13–7.
4. Breiman L. Statistical modeling: two cultures. Stat Sci. 2001;16:199–215.
5. Burkov A. The hundred-page machine learning book. Andriy Burkov. 2019.
6. Donoho D. 50 years of data science. J Comput Graph Stat. 2017;26:745–66.
7. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of p-hacking in science. PLoS Biol. 2015;13(3):e1002106. https://doi.org/10.1371/journal.pbio.1002106.
8. Li JJ, Tong X. Statistical hypothesis testing versus machine learning binary classification: distinctions and guidelines. Patterns. 2020;1:100115.
9. Shane J. You look like a thing and i love you: how artificial intelligence works and why it's making the world a Weirder place. New York: Little, Brown, and Company; 2019.
10. Vollmer S, Mateen BA, Bohner G, Kiràly FJ, Ghani R, Jonsson P, Cumbers S, Jonas A, McAllister KSL, Myles P, Grainger D, Birse M, Branson R, Moons KGM, Collins GS, Ioannidis JPA, Holmes C, Hemingway H. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. Br Med J. 2020;368:169297. https://doi.org/10.1136/bmj.16927.
11. Waller LA, Miller GW. More than manuscripts: reproducibility, rigor, and research productivity in the big data era. Toxicol Sci. 2016;149:275–6. https://doi.org/10.1093/toxsci/kfv330.
12. Waller L, Levi T. Building intuition regarding the statistical behavior of mass medical testing programs. Harvard Data Science Review; 2021. Retrieved from https://hdsr.mitpress.mit.edu/pub/hodep31o.
13. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appelton G, Azton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roose M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenberg P, Wolstencroft K, Zhao J, Mons B. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.
14. Yoder RM. Someday they'll get slick Willie Sutton. The Saturday Evening Post. 1951;223(30). Saturday Evening Post Society, Indianapolis.