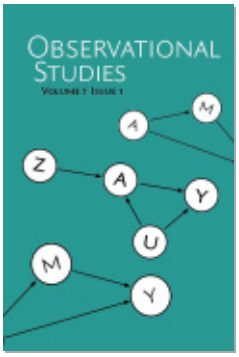
Beyond Two Cultures: Cultural Infrastructure for Data-driven Decision Support



Nikki L. B. Freeman, John Sperger, Helal El-Zaatari, Anna R. Kahkoska, Minxin Lu, Michael Valancius, Arti V. Virkud, Tarek M. Zikry, Michael R. Kosorok

Observational Studies, Volume 7, Issue 1, 2021, pp. 77-94 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2021.0024*

Observational Studies Vol. 7.1 (2021): Published July 2021

Beyond Two Cultures: Cultural Infrastructure for Data-driven Decision Support

Nikki L. B. Freeman[1] **nlbf@live.unc.edu**
Department of Biostatistics
University of North Carolina at Chapel Hill

John Sperger[1] **jsperger@live.unc.edu**
Department of Biostatistics
University of North Carolina at Chapel Hill

Helal El-Zaatari **helal@live.unc.edu**
Department of Biostatistics
University of North Carolina at Chapel Hill

Anna R. Kahkoska **anna_kahkoska@med.unc.edu**
Department of Nutrition
University of North Carolina School of Medicine

Minxin Lu **mino12@live.unc.edu**
Department of Biostatistics
University of North Carolina at Chapel Hill

Michael Valancius **mval@email.unc.edu**
Department of Biostatistics
University of North Carolina at Chapel Hill

Arti V. Virkud **avirkud@unc.edu**
Department of Epidemiology
University of North Carolina at Chapel Hill

Tarek M. Zikry **tarek@live.unc.edu**
Department of Biostatistics
University of North Carolina at Chapel Hill

Michael R. Kosorok **kosorok@live.unc.edu**
Department of Biostatistics
University of North Carolina at Chapel Hill

Abstract

In the twenty years since Dr. Leo Breiman's incendiary paper *Statistical Modeling: The Two Cultures* was first published, algorithmic modeling techniques have gone from controversial to commonplace in the statistical community. While the widespread adoption of these methods as part of the contemporary statistician's toolkit is a testament to Dr. Breiman's vision, the number of high-profile failures of algorithmic models suggests that Dr. Breiman's final remark that "the emphasis needs to be on the problem and the data" has been less widely heeded. In the spirit of Dr. Breiman, we detail an emerging research community in statistics — data-driven decision support. We assert that to realize the full

1. Equal contribution, co-first authors

potential of decision support, broadly and in the context of precision health, will require a culture of social awareness and accountability, in addition to ongoing attention towards complex technical challenges.

Keywords: Precision health, Decision Support, Machine Learning, Two Cultures

1. Introduction

As we mark the twentieth year anniversary of Dr. Leo Breiman's *Two Cultures* (Breiman et al., 2001), it is clear that the statistics cultural landscape has and continues to be shaped by Dr. Breiman's writing. There is a reason why 'two cultures' is a shorthand for differing meta-statistical frameworks for the practice of statistics. Right or wrong, in full agreement, half agreement, or outright rejection of Dr. Breiman's claims, the bold contrasts that Dr. Breiman delineated between what he labeled the data modeling and algorithmic modeling cultures, the deep assumptions of each culture, and the consequences of those cultures for the future are, in our opinion, what continues to attract interest and stimulate spirited discussion of his work.

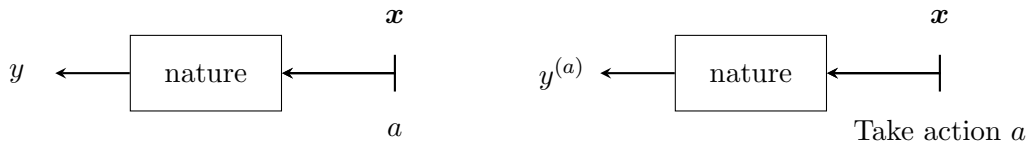
It is undeniable that Dr. Breiman was writing in a very different statistical landscape than the one that statisticians encounter today. Among many changes since the publication of *Two Cultures*, we have experienced the rise of "big data" and the technological advances in collecting, storing, and processing those data (Hilbert and Lopez, 2011). Massive advances in computing hardware and software and accompanying theoretical advancements have accelerated developments in computational statistics; in parallel, advances have come in the application and development of statistical methodology to fields such as genetics and neuroscience (Bielza and Larranaga, 2020; Balding and Bishop, 2001; Bottou et al., 2018; Gentle et al., 2012). Beyond the types of data routinely used and the technologies to analyze it, the statistician's toolbox has changed too; machine learning, along with continued development of its theoretical underpinnings and cultural acceptance, is now a vital statistical approach (Hastie and Tibshirani, 2009). Perhaps the most pertinent to *Two Cultures* is the work that has been done to mediate the trade-offs that Dr. Breiman described, modeling that is both interpretable and accurate (Rudin, 2019).

The very future we can imagine for statistics has changed too. We have previously argued that data-driven decision support is an exciting frontier for statisticians and the field at large (Sperger et al., 2020). Data-driven decision support is fundamentally different from the tasks of prediction and attribution that characterized the world that Dr. Breiman described. As we detail later, data-driven decision support modeling is causal, dynamic and potentially continuously learning, and decision support tools interface with and impact society in profound ways. In the spirit of Dr. Breiman, we look to this future. We articulate the characteristics of the statistical culture we hope to cultivate for the decision support research community, a culture that aims for social awareness and accountability. We focus on decision support in the context of precision health, a paradigm for individualized treatment selection, but our observations and call to action apply to decision support contexts more broadly. We begin by describing statistical decision support and offering concrete examples of current work in the field. We then highlight two cultural elements that we hope will characterize the statistical culture around decision-support and reflect on the evolving role of statisticians in the decision support research community.

2. Data-driven decision support

In *Two Cultures*, Dr. Breiman delineated two goals of regression: prediction and [the extraction of] information. Dr. Bradley Efron, reflecting on Dr. Breiman, identified the three fundamental aims of regression: prediction, estimation, and attribution (Efron, 2020). In light of our own work in precision health, we suggest an addition to Dr. Efron's list: decision-making or decision support.

Decision-making is a fundamentally causal exercise, requiring us to consider the "what-if" scenarios that could result from our actions. Dr. Breiman concisely visualized regression data as a vector of inputs that is transformed by nature into a response. Building on Dr. Breiman's abstraction, we can think of the data generated from a decision-making problem as the result of nature taking a vector of input variables $\boldsymbol{x}$ and an action a to produce a response y (Figure 1). Breaking the action a out from the other input variables $\boldsymbol{x}$ emphasizes what distinguishes causal inference and formal statistical decision support from other statistical tasks — a is manipulable. Breaking out a also changes how we think about the response. Often formalized through potential outcomes (Rubin, 1978, 2005; Hernán and Robins, 2010), the responses we consider are what we would observe if we were to take action a and denote the potential response for a particular action a as $y^{(a)}$. Of course, potential outcomes are precisely that, potential. While we have the potential to observe any one of the potential outcomes before an action is taken, only one action can be taken and thus at best only one potential outcome can be observed.



(a) Observed (b) Potential

Figure 1: Data Generation

One compelling statistical formulation of data-driven decision support comes from the precision health literature. In the field of health sciences and public health research, precision health is a shift away from the traditional one-size-fits-all approach of comparing averages across individuals who did or did not receive a treatment and seeks to leverage patient heterogeneity in a causally valid framework by using data to determine which patients should receive which treatment to optimize key health outcomes (Kosorok and Laber, 2019). It is formalized through dynamic treatment regimes (DTRs), a sequence of decision rules, one for each key decision point, that maps patient features to a treatment recommendation; an optimal dynamic treatment regime is a DTR that optimizes patient outcomes on average across the target population. Dynamic in this context can refer to differences among patients, decisions that take place sequentially over time, or both. DTRs are often learned through reinforcement learning. One special benefit of the reinforcement learning approach is the ability to account for the long term effects of treatments and synergies between treatment sequences while learning optimal DTRs. Depending on the precision health problem, reinforcement learning may be done online or offline and the problem may be finite horizon

Research Question	Statistical Goal
What is Patient X's 5-year survival probability based on their individual characteristics (including age at diagnosis) and current treatment regimen?	Prediction
How does age at diagnosis impact 5-year survival across the population of patients with lung cancer?	Attribution
What treatment regimen (e.g. chemotherapy versus radiation therapy) should Patient X receive to maximize likelihood of 5-year survival, based on their individual characteristics?	Decision Support

Table 1: Statistical goals and illustrative research questions

or infinite horizon. The resulting optimal DTR can be used to help individuals, medical providers, and communities make decisions that are tailored to individualized health care needs.

Considering a few simplified research questions will help illuminate the distinction between decision support and other statistical goals, and Table 1 provides examples of research problems tied to the goals of prediction, attribution, and decision support. When the goal is prediction, there are potentially many factors that may improve prediction and none of them need to be factors that we think are causal in nature. When the goal is attribution, such as estimating the effect of age at cancer diagnosis on 5-year survival, the problem may sound like a causal problem but is not; age is a non-modifiable risk factor from an interventional perspective. While these prediction and attribution research questions can be potentially practice changing questions, they fall outside the purview of decision support. In contrast to prediction and attribution, consider the question of which treatment(s) to start for an individual lung cancer patient based on their clinical characteristics such as age, cancer stage, and specific genetic or phenotypic biomarkers. For this decision-making problem, there is an embedded causal question: *what would happen under each potential treatment regimen?* This question inherently recognizes the existence of heterogeneity between patients and rather than seeking to answer "Which treatment?" seeks to answer "Which treatment for whom?". Notably, this question also suggests another distinction: predicting the response is not necessary for decision-making (though it is often of separate scientific interest); estimating the difference in outcome under the potential interventions while treating the variables that affect the outcome as nuisance parameters is enough provided the outcome appropriately incorporates benefits and costs. This insight has been leveraged by methods like advantage learning (Murphy, 2003).

Statistical decision support, as we have described it, can manifest as tools for augmenting human decision-making or fully-automated decision-making systems. This covers a wide range of tools from simple checklists to automated insulin delivery systems, better known colloquially as the "artificial pancreas" (Dassau et al., 2013). We will take an expansive view of decision support to include both types of tools; while they vary widely in complexity, in each case their success depends on attending to the context and consequences of the decision in addition to statistical concerns. For this reason, we believe it is important to explicitly define the challenges of decision support and name the cultural elements that must

be cultivated to address those challenges; this exercise may guide future work in decision support research.

3. Real-world examples of decision support

Using observations from his own practice of statistics in industry and academia, Dr. Breiman described the 'two cultures' of statistical modeling. Like Dr. Breiman, we will use examples from our research group to illustrate the challenges of statistical decision support. Our hope is that these examples will convey the variety of applied research problems in decision support and highlight the most salient considerations for the emerging decision support research community.

3.1 Type 1 Diabetes

Type 1 diabetes (T1D) is an autoimmune disease whereby the body attacks and destroys its insulin-producing beta cells, resulting in lifelong insulin dependency and a demanding self-management routine (Daneman, 2006). The prevalence of obesity and overweightness in individuals with T1D has recently increased and now parallels that of the general population (Liu et al., 2010; DuBose et al., 2015). Unfortunately, T1D is associated with a greater risk of cardiovascular disease and excess adiposity futher compounds this risk (Corbin et al., 2018).

The management of T1D itself can be a barrier to exercise and weight loss due to challenges with blood glucose management including the increased risk of and fear of potentially-deadly episodes of hypoglycemia, the state of having low blood sugar and needing rapid carbohydrate consumption (Zaharieva et al., 2020). Although there are clinical guidelines for blood glucose management around exercise (Riddell et al., 2017), significant inter- and intra-individual heterogeneity exists, complicating management on a day-to-day or even minute-to-minute basis (Riddell et al., 2019; Zhao et al., 2013). As a result, individuals with type 1 diabetes must try to accurately estimate insulin needs and dosing in consideration of future activities such as exercise, due to the delayed effect of exogenous insulin and the body's rapidly changing glucose needs during and after physical activity.

Suppressing the dependence on time and returning to our notation from the previous section, we now describe how we formalized the T1D decision problem. The available predictors $\boldsymbol{x}$ included the history of blood glucose level (real-time using continuous glucose monitoring), physical activity (real-time based on accelerometer data), and insulin doses logged by an insulin pump. The actions a that could be recommended included all possible combinations of food consumption, additional boluses of insulin, increase or decrease in insulin doses, increase in physical activity, and no action at this time. The outcome y was a severity-weighted sum of glycemic events over the 60 minutes preceding and following time t that captures excursions outside the normal blood glucose range aligned with their clinical importance and risk to the individual.

Using data from a pilot study with 31 adolescent patients with type 1 diabetes to address the need for individualized, real-time strategies, we developed a decision support tool based on a Markov decision process and estimated dynamic treatment regimes using a V-learning approach that models the minute-to-minute (infinite horizon) decision-making required by diabetes management (Luckett et al., 2019). V-learning models the state-value function;

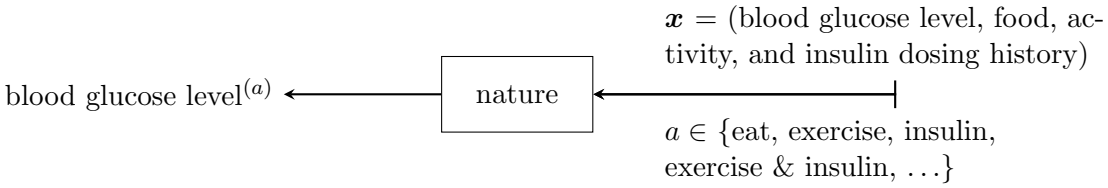


Figure 2: T1D Data

linear, polynomial, and Gaussian basis function approximations were used with different discount factors to find the optimal model.

The algorithm is currently being developed into a mobile health (m-health) application that will collect an individual's history of food intake, exercise, and insulin dosing to provide individual-specific real-time recommendations. The application under development generates real-time decision support recommendations that are highly tuned to individual goals, physiology, and real time data, the goal of which is to offer exercise autonomy for people with diabetes by utilizing data already collected as part of their existing care. The process of developing the tool has and continues to demonstrate what makes decision support unique from other statistical tasks. For example, the question of what should be included directly into the optimization problem versus accounted for qualitatively was a critical question in the development process. The tool optimizes for glycemic control, but other objectives such as weight loss were considered and may be of interest to integrate later. Moreover, because the recommendations to users are not mediated through a health care professional, interpretability, trust, and acceptability of the tool continue to be important guideposts in the development process, necessitating waves of end-user and other stakeholder engagement.

3.2 Knee Osteoarthritis

Knee osteoarthritis (KOA) is one of the leading causes of pain and disability in adults worldwide (Cross et al., 2014). Weight loss and exercise are known to reduce pain and improve function, but there is substantial heterogeneity in the amount of benefit; there are concerns that patients who are extremely overweight may be harmed by exercise until they bring their weight below a certain threshold. Identifying who to recommend diet alone, exercise alone, or diet & exercise to may provide clinical benefit to KOA patients over a blanket recommendation of diet & exercise.

The Intensive Diet and Exercise for Arthritis (IDEA) trial randomized 343 participants to one of three interventions: exercise alone, diet alone, or diet & exercise to determine whether diet & exercise was more effective than diet alone or exercise alone for treating KOA patients (Messier et al., 2013). While it was not designed with decision support in mind, the trial's rich data set of baseline covariates and outcome data, multiple actions, and suspected heterogeneity in treatment response made the study a promising candidate for applying precision health thinking and techniques. In this light, the challenge is to identify who should be recommended exercise alone, diet alone, and diet and exercise in combination based on a set of 76 baseline covariates $\boldsymbol{x}$ that included demographic and clinical information with measures of pain, functioning, stiffness, and physical activity. The situation was complicated by seven relevant outcomes including pain, function, a biomarker Interleukin-

6, and stiffness. These outcomes may involve trade-offs; for example, a treatment that decreases pain could do so at the expense of functioning. One could choose to try to improve a single outcome or use methods that can accommodate multiple outcomes.

A variety of methods were compared including penalized regression methods, list-based methods, random forests, and deep learning for single and multiple outcome cases. For weight loss, the estimated optimal decision rule assigned diet alone to patients with a baseline weight above 109 kilograms and a waist circumference above 90 centimeters while everyone else was assigned to diet and exercise. One exception was made to this rule: everyone with a history of heart attack was assigned to diet and exercise. The result was a modest clinical improvement in weight loss of 1.4 kilograms at 18 months that was statistically significant ($p = .01$) (Jiang et al., 2020). While this level of improvement may be relatively modest, the simplicity of the treatment rule would make it easy to implement in clinical practice, and it could have an important impact at the population level.

Because of the nature of the data available, the action set considered was limited to exercise alone, diet alone, and exercise and diet in combination. In a more general analysis, it may be desirable to consider a larger action set. Here we see the impact of data realities and early data collection decisions on decision support. Future work on this problem may also want to include patient preferences for outcomes and the feasibility of the interventions across the heterogeneous patient population. Still, this precision health problem showcases the promise of individualized decision support and points to how to improve the way we approach decision support research.

4. Towards a unified culture for decision support

While we are excited about statistical decision support and its potential for improving human lives, we are fully cognizant of the challenges it presents and the potential for unintentional harm. A number of examples of decision support gone awry, ranging from Amazon's failed attempt to automate hiring decisions to a racially biased risk calculator used to identify patients with complex health needs, can be found in academic literature and the popular press (Dastin, 2018; Obermeyer et al., 2019). Fortunately, the emerging community of researchers around statistical decision support and its application areas have recognized the potential for decision support tools to cause unintended consequences (Cathy O'Neil, 2016; Zarsky, 2016; Doshi-Velez and Kim, 2017; Crawford, 2017).

At the core of the 'two cultures' that Dr Brieman described was a difference in fundamental assumptions and approaches to statistical problems, and how the consequences of those assumptions are borne out in statistical modeling. In the context of decision-support, we too believe that researchers' beliefs and their manifestation in the scientific process directly influence the quality and impact of the resulting work. Grounded in our experiences and the thoughtful reflections of the research community engaging in decision support research, we endeavor to characterize two cultural elements that we believe the decision support research community has begun to and should continue to nurture:

- Social awareness — approach statistical modeling with an eye to fairness, interpretability, and understanding of the context in which the decision support will be employed

- Accountability — be transparent during all steps of research, and develop tools and procedures so that decisions can be justified and their provenance known

Unlike Dr. Breiman, whose discussion was framed as trade-offs and inherent tensions between the 'two cultures', in the following, we frame our discussion in the need for and importance of a unified, forward-facing culture for decision support modeling. We also envision an expanded role for statisticians as a bridge in this culture between stakeholders and the decision support development process.

4.1 A culture of social awareness

We believe that applying decision support tools to real world decisions should bring us closer to an ideal world, not farther. Doing this requires a deep understanding of and accounting for the context in which decisions are made and the broad implications of those consequences throughout the development process, including statistical modeling. A number of conceptual models from a variety of fields and with various use-cases in mind have been developed to represent the wide perspective needed when thinking about decision making and its consequences. For example, the socio-ecological model of health conceptualizes health as a result of a combination of and interplay between intra- and interpersonal, institutional, community, and policy factors (Bronfenbrenner, 1977, 1992). Extending this rationale to decision support, any decision support tool implemented necessarily interacts, affects, and is affected by each level highlighted in the socio-ecological model. For decision support tool developers, this means that modeling considerations and assessments of performance should proactively take in account end users and their communities as well as society as a whole. More recent work has also called for the use of multidisciplinary complex adaptive system theory and other system dynamics-based methods to model complex societal context for machine learning problems (Martin Jr et al., 2020b).

Designing decision support with this perspective is a broad charge. We do not contend to have an optimal or universal algorithm to recommend how to develop socially conscious decision support tools. Nonetheless, we do believe that good decision support begins with social awareness and that prioritizing the context and consequences of decisions throughout the decision support development and implementation process is essential for the future success of the field and for decision support to reach its potential at improving human lives (Robinson et al., 2019). Consider the decision support development life-cycle: it demands data collection, storage, model or tool development, and cross-disciplinary engagement for implementation and on-going evaluation. At each step, because of structural barriers, systemic biases, barriers to change, cost constraints, and plain unintentional mistakes, disadvantaged groups are at greater risk of harm. Because of expense or difficulty in sampling, they may be under-represented in data collection. Upstream under-representation has downstream implications for the decision support performance and some groups may disproportionately bear the burdens of bad decision support. Failure to include a diverse, representative group of stakeholders can yield decision support that after a long development process is not acceptable to certain communities. Moreover, some groups may not have a voice in whether they are subjected to interventions aided by a decision support tool, may not have a say in what outcomes are maximized, or may not bear the benefit of decision support because it is not implemented in their communities. For all of these reasons, what

we are calling social awareness–engaging with stakeholders throughout the entire decision support development process and integrating context-specific needs into decision support tools–is essential for giving decision support developers the best chance at avoiding social harm, identifying and fixing problematic decision support tools early in the process, and building trust with decision-makers and their communities.

As statisticians, we believe that a starting place for bringing social awareness to statistical practice is interpretable modeling and fairness. We start with interpretability. An interpretable model is desirable not only because it can be easier to explain to and be accepted by users (Rudin and Ustun, 2018) but also because it is easier for users to remember when they apply the knowledge in their practice and can provide insights for future studies (Zhang et al., 2018). Although there is no universal mathematical definition of interpretability, a deep literature exists that explores model interpretability, model explanability, and real and perceived tensions with model accuracy (Rudin, 2019; Ribeiro et al., 2016; Lundberg and Lee, 2017). In the context of decision support modeling, the interpretability of a model is not decided by developers but by users. Notably, the background of decision support tool developers and the background of users can be very distinct; what is interpretable to a model developer may not be interpretable to the end user. Making decision recommendations and the processes by which those recommendations are made interpretable is important for building trust with end users, especially for those users who have been harmed historically by less socially-grounded algorithmic decision tools. Incomplete, misleading, or non-transparent information can lead to faulty decision support and can harm decision makers and the decision support field severely. Efforts have been made in the field to build models that are both accurate and interpretable in more general settings, such as interpretable dynamic treatment regimes, which produces a simple rule as a list of "if. . . then. . . " statements (Zhang et al., 2018); and Risk-Calibrated Supersparse Linear Integer Models (RiskSLIM), which perform as effectively as the best black-box algorithm on recidivism prediction (Rudin and Ustun, 2018).

Closely related to interpretable models is the notion of fairness. While interpretability is concerned with the form of the model be it a decision list or a more complicated form, fairness relates to the decision rules learned and outputted by a model. Interpretability may help us assess why or why not a model is or is not fair, but interpretability does not guarantee fairness. Multiple definitions of fairness have been offered to different aspects of fairness in statistical modeling. Corbett-Davies and Goel (2018), in addition to offering their own standard of fairness, delineated between different definitions of fairness in machine learning and showed that each definition has drawbacks and how striving for fairness by one definition may yield discriminatory results by another definition. Despite the difficulties in defining objectively fair models, model transparency and deep interrogation of decision support outputs can help ensure that decision support tools do not worsen existing structural inequalities among those affected by algorithmic decision aids and tools. To this end, new areas of research are exploring approaches to ensure stakeholder-engaged problem representation and mitigate bias in algorithm development and deployment, including community based system dynamics (Martin Jr et al., 2020b,a).

Interpretability and fairness are only a starting place for socially aware, context-informed decision support. We also recognize that sometimes interpretability will need to be sacrificed to better performing but less interpretable black box-like approaches, at least temporarily,

to ensure optimal (health) outcomes, but fairness should never be sacrificed. Integrating interpretability and fairness together into modeling along with concerns from the level of the decision maker to the level of society and the environment are important future areas of exploration for statisticians and their collaborators engaged in decision support research.

4.2 A culture of accountability

Statisticians working in decision support must cultivate responsibility for both the technical aspects of their work and, as the preceding section has hopefully made clear, the wider social implications of their work. This kind of responsibility is not new in statistics, and broad guidelines have been codified by professional organizations like the American Statistical Association and the Royal Statistical Society (Committee on Professional Ethics of the American Statistical Association, 2018; Royal Statistical Society, 2014), but the immediacy of decision support requires greater transparency, community involvement, and oversight compared to other applied statistical endeavors. Nowhere is this more important than m-health, and consumer-directed mobile applications generally, because the recommendations are not filtered through an expert intermediary such as a physician. Accountability extends beyond meeting legal requirements to ensuring that actions are justifiable and fair and includes being able to track the provenance of a decision recommendation. We will first review the challenges of accountability with decision support and discuss how transparency will be key to addressing these challenges. Afterwards, we will look at how these considerations are playing out in the context of health care in the United States as a reference model. While every industry will need to define its own standards for accountability commensurate to the impact of the decisions made in the field based on these principles, the high-stakes nature of the decisions in medicine have already prompted regulatory changes in the US through the 21st Century Cures Act as well as international recognition that regulatory agencies need to continue developing new approaches to meet the evolving challenges of decision support.

Evolving decision support tools, especially those involving automated decision-making, present an emerging challenge for both governmental and study-level oversight. While decision support tools have existed for decades, the FDA has recognized the need for modernizing its approach to regulating software-based decision support systems and recently issued draft guidance about its planned approach to regulating clinical decision support systems (FDA, 2019a). Decision support systems can evolve over time as they accumulate data, and their performance can change over time even if the tool remains static. Concept drift, in which previous data is no longer representative of new data, can render a batch-learned model biased. Though online learning methods that continuously update the model as data accumulates can remedy this issue, they present their own challenges. In online learning, the decision-making algorithm can evolve without a single line of code changing in the software and the performance and sensibility of the model must be regularly evaluated. The ability to trace the provenance of a recommendation will be critical for navigating these challenges, but as Dr. Michael Jordan (Jordan, 2019) noted recently, trying to track down the provenance of many medical recommendations is challenging for researchers, let alone the general public. Other challenges particularly salient in decision support include the extent to which a decision is explainable and whether the tool is directed at patients or health care providers.

While regulation is important, accountability is about more than meeting governmental regulations. Transparency plays a fundamental role in enabling accountability, and this transparency must cover all aspects of the model development and deployment process. In practice this means that the reasons for a decision should be communicated at an appropriate level of detail, and that the underpinnings should be both explainable and traceable, including: what data was used to train the model and how was it collected, what inputs does the model use and how does it weight them, how well does the model perform, what was explicitly optimized, and what important considerations may have been left out of the optimization and left as qualitative considerations. We will focus on some of the challenges that are unique or heightened for decision support, but will quickly note that many of the concerns with modern statistical and machine learning applications like privacy (Harman et al., 2012; Papernot et al., 2016) that have been covered extensively elsewhere are also concerns here. In a decision support analysis, the set of potential actions and the outcome(s) of interest must be explicitly specified, and rarely are all of the relevant outcomes or potential actions included directly in an optimization problem for decision support. By itself, this is not necessarily an issue; many considerations may be more amenable to qualitative or individual consideration than formal optimization, and certain actions and data may be too costly or involve ethical concerns. The choice of what to include can become problematic if there is not clear communication about what the decision-making algorithm considers and what important factors are omitted, or if the exclusions are motivated by ethically or scientifically wrong reasons. We are also limited in what is under our control, while an outcome of interest may be effected by actions from many actors from individuals to doctors to families and peer groups all the way up to the government. The choice of model is typically under our control as statisticians, and whether black-box methods should be used for high-stakes decisions is being contentiously debated (Rudin, 2019).

To see how the issues of explainability, fairness, and ongoing monitoring are playing out in a regulatory setting, we will briefly look at how the FDA is approaching these issues. The FDA's draft guidance on clinical decision support software considers both the intended audience (physician or patient) and the level of explainability in the determination of whether the application is tightly regulated, technically regulated but currently not enforced at the agency's discretion, or whether it is outside of the agency's purview (FDA, 2019a). The FDA does not regulate physician-targeted tools when the "logic and inputs" are available to the physician, while for patient-targeted tools the decision of whether they plan to enforce compliance depends on the risk-level of the condition. Maintaining a bright line between what is explainable and what is not may prove challenging (Evans and Ossorio, 2018), and we hope that the FDA emphasizes that interpretability is ultimately an empirical question about how physicians use and understand it, not a fundamental property of a model. Regardless of whether the FDA changes tack in future guidance, researchers involved in developing decision support systems must carefully investigate whether their tools are actually correctly understood by the people using them and not be content with simply having applied an interpretable method.

The FDA has also begun to examine how other big data sources like electronic health record and claims data (linked and unlinked) can be used in decision support. In doing so they have started to build a framework for "real world evidence"(RWE), a newly coined termed representing evidence outside of clinical trials (Schurman, 2019). While clinical

decision support systems are not mentioned explicitly in connection with their RWE program, this kind of evidence may be critical for both the FDA and researchers to evaluate decision support systems in a way that better reflects their usage in practice. In contrast to the explainability consideration, the draft guidance does not explicitly address fairness or ongoing monitoring. Instead, the FDA requires that developers create requirements for continued monitoring that their products are safe and effective because quality standards are context-specific (FDA, 2019b). In the context of decision support, it is clear that monitoring both ongoing performance and fairness should be part of quality standards even when the method is not explicitly considering subgroups.

4.3 Final reflections from real-world examples

Our experience with the T1D m-health app and its ongoing development has required us to navigate social awareness and accountability to multiple stakeholders in decision support research. Accountability to patients is both a priority and an ongoing process. Our first goal was to elicit and explore the needs of patients by undertaking a qualitative study to characterize the patient-perceived experience and barriers to weight management among youth and adolescents, including exercise (Kahkoska et al., 2018). Although individuals in the study indicated that they were already adjusting glucose management behaviors when exercising, they reported frustration with differences in day-to-day outcomes and reported struggling with episodes of high and low blood sugar despite repeating strategies that had previously worked. Youth in the study explicitly expressed a desire for individualized recommendations to avoid low or high blood sugar levels, including before, during, and after exercise sessions. These findings were then replicated in a young adult population (Addala et al., 2019). Together, the qualitative data were formative in conceiving the decision support tool; an understanding of lived patient experiences revealed that a real-time, individualized recommendation to optimize blood glucose around planned exercise could help to reduce the risk of exercise-induced hypoglycemia and facilitate safe and effective exercise for this population, thereby reducing cardiovascular risk, improving weight status, and promoting overall physical and mental well-being.

Development of the m-health app is ongoing, and accountability continues to be a key consideration as we move forward. Additional studies will be needed to engage individuals with type 1 diabetes and improve the m-health system in terms of future usability and performance, revise or develop additional functionality, and further improve the user interface. Once acceptability has been determined, there are additional, critical steps required to validate the safety of the model's patient-specific analysis and treatment recommendations, requiring input from endocrinologists and other care providers, as well as regulatory approval by the FDA and ongoing evaluation. Sharing the tool across academic and industry communities may generate new feedback to further improve the tool. Specific and thorough training materials will be necessary to ensure that users can successfully on-board and understand how to safely use the tool, including reconciling strategies against their personal judgement or medical advice when it may conflict.

5. Conclusion

Dr. Breiman framed the goal of statistics as "us[ing] data to solve problems". The decision support problems we have described through our examples and exposition are complicated, but they represent direct applications of statistics to solve problems. Decision support warrants both a broad and deep understanding of the context in which the decision support tool will be deployed, including outcomes of interest, their underlying determinants, related outcomes, the mediators of changes in outcomes, and the potential for unanticipated effects. The potential impacts of algorithms directly recommending human actors to take actions they may not otherwise take requires strategies for oversight and accountability that evolve in tandem with the field. Moreover, decision support modeling can not only propagate but intensify existing disparities if care is not taken to identify where these patterns may occur.

These challenges, coupled with the complexity of each stage of developing and launching a decision support tool, ensure no one individual can bring all of the expertise and insight required. We believe that decision support modeling should be considered a team activity grounded in a scientific culture characterized by, but perhaps not exclusively by, the common goals of social awareness and accountability. On these teams, we foresee a new, evolving role for statisticians as the hub for collaboration. From this perspective, statisticians are the central translator of personal, community, social, and environmental insights into concrete model choices. Statistical training will need to reflect this shift, emphasizing both the deep technical skills required to solve decision support problems as well as the leadership and communication skills needed to excel in this role.

Assumptions about model form, as Dr. Breiman so clearly pointed out, are manifestations of our beliefs about the world and how the data we learn from are generated. Concretely for statisticians in the decision support field, the act of translating beliefs to models will include mapping domain knowledge and multiple simultaneous social imperatives to objective functions that reward for good decisions, capture contextual nuance among heterogeneous decision makers, and budget for real world constraints. It will include approaching model selection with technical performance metrics as well as metrics of social fairness and community specific needs. Throughout, interpretability and transparency in the modeling development and evaluation process will be key. Inevitably, these models will need to be constructed, scrutinized, and constructed again.

In the spirit of Dr. Breiman, we believe that statistical culture is crucial infrastructure for the future of the field. On a cultural foundation of social awareness and accountability, we believe that data-driven decision support can be elevated to reach its highest potential.

Acknowledgments

Dr. Kosorok acknowledges support from the National Cancer Institute under grant award P01 CA142538. Tarek M. Zikry acknowledges support from the NHLBI under NIH award 1F31HL156464-01. Dr. Kahkoska is supported by the National Institute Of Diabetes And Digestive And Kidney Diseases of the National Institutes of Health under Award Number F30DK113728. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute, the National Na-

tional Heart, Lung, and Blood Institute, National Institute Of Diabetes And Digestive And Kidney Diseases, or the National Institutes of Health.

References

21st Century Cures Act. H.R. 34, 114th Cong. (2015).

Ananta Addala, Daria Igudesman, Anna R Kahkoska, Franklin R Muntis, Katherine J Souris, Keri J Whitaker, Richard E Pratley, and Elizabeth Mayer-Davis. The interplay of type 1 diabetes and weight management: A qualitative study exploring thematic progression from adolescence to young adulthood. Pediatric diabetes, 20(7):974–985, 2019.

D. J. Balding and M. Bishop, editors. Handbook of statistical genetics. Chichester ; New York : Wiley, 2001.

Concha Bielza and Pedro Larranaga. Data-driven computational neuroscience : machine learning and statistical models. Cambridge, United Kingdom ; New York, NY : Cambridge University Press, 2020.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization Methods for Large-Scale Machine Learning. SIAM Review, 60(2):223–311, January 2018. doi: 10.1137/16M1080173.

Leo Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science, 16(3):199–231, 2001.

Urie Bronfenbrenner. Toward an experimental ecology of human development. American Psychologist, 32(7):513–531, July 1977. ISSN 0003-066X. doi: 10.1037/0003-066X.32.7.513.

Urie Bronfenbrenner. Ecological systems theory. In Six theories of child development: Revised formulations and current issues., pages 187–249. Jessica Kingsley Publishers, London, England, 1992. ISBN 1-85302-137-7 (Paperback).

Cathy O'Neil. Weapons of Math Destruction : How Big Data Increases Inequality and Threatens Democracy. Crown, New York, 2016. ISBN 978-0-553-41881-1.

Committee on Professional Ethics of the American Statistical Association. American statistical association ethical guidelines for statistical practice, 2018. URL https://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx. Accessed: 2021-02-26.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023, 2018.

Karen D Corbin, Kimberly A Driscoll, Richard E Pratley, Steven R Smith, David M Maahs, Elizabeth J Mayer-Davis, Advancing Care for Type 1 Diabetes, and Obesity Network (ACT1ON). Obesity in type 1 diabetes: pathophysiology, clinical impact, and mechanisms. Endocrine reviews, 39(5):629–663, 2018.

Kate Crawford. The trouble with bias, December 2017. URL https://nips.cc/Conferences/2017. Invited talk at Thirty-first Conference on Neural Information Processing Systems.

Marita Cross, Emma Smith, Damian Hoy, Sandra Nolte, Ilana Ackerman, Marlene Fransen, Lisa Bridgett, Sean Williams, Francis Guillemin, Catherine L Hill, et al. The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study. Annals of the rheumatic diseases, 73(7):1323–1330, 2014.

Denis Daneman. Type 1 diabetes. The Lancet, 367(9513):847–858, 2006.

Eyal Dassau, Howard Zisser, Rebecca A Harvey, Matthew W Percival, Benyamin Grosman, Wendy Bevier, Eran Atlas, Shahar Miller, Revital Nimri, Lois Jovanovič, et al. Clinical evaluation of a personalized artificial pancreas. Diabetes care, 36(4):801–809, 2013.

Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, October 2018. URL https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. February 2017. URL http://arxiv.org/abs/1702.08608v2.

Stephanie N DuBose, Julia M Hermann, William V Tamborlane, Roy W Beck, Axel Dost, Linda A DiMeglio, Karl Otfried Schwab, Reinhard W Holl, Sabine E Hofer, David M Maahs, et al. Obesity in youth with type 1 diabetes in germany, austria, and the united states. The Journal of pediatrics, 167(3):627–632, 2015.

Bradley Efron. Prediction, estimation, and attribution. International Statistical Review, 88:S28–S59, 2020.

Barbara Evans and Pilar Ossorio. The challenge of regulating clinical decision support software after 21st century cures. American journal of law & medicine, 44(2-3):237–251, 2018.

FDA. Clinical Decision Support Software — Draft Guidance for Industry and Food and Drug Administration Staff. U.S. Department of Health and Human Services Food and Drug Administration, September 2019a. URL https://www.fda.gov/media/109618/download.

FDA. Policy for Device Software Functions and Mobile Medical Applications. U.S. Department of Health and Human Services Food and Drug Administration, September 2019b. URL https://www.fda.gov/media/80958/download.

James E. Gentle, Wolfgang Karl Härdle, and Yuichi Mori. How computational statistics became the backbone of modern data science. In James E. Gentle, Wolfgang Karl Härdle, and Yuichi Mori, editors, Handbook of Computational Statistics. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-21550-6 978-3-642-21551-3. doi: 10.1007/978-3-642-21551-3.

Laurinda B Harman, Cathy A Flite, and Kesa Bond. Electronic health records: privacy, confidentiality, and security. AMA Journal of Ethics, 14(9):712–719, 2012.

Trevor Hastie and Robert Tibshirani. The elements of statistical learning : data mining, inference, and prediction. Springer series in statistics. New York : Springer, second edition, corrected 7th printing. edition, 2009. ISBN 978-0-387-84858-7 978-1-282-12674-9 978-0-387-84857-0 978-0-387-84884-6.

Miguel A Hernán and James M Robins. Causal inference, 2010.

Martin Hilbert and Priscila Lopez. The World's Technological Capacity to Store, Communicate, and Compute Information. Science, 332(6025):60–65, 2011.

Xiaotong Jiang, Amanda E Nelson, Rebecca J Cleveland, Daniel P Beavers, Todd A Schwartz, Liubov Arbeeva, Carolina Alvarez, Leigh F Callahan, Stephen Messier, Richard Loeser, et al. A precision medicine approach to develop and internally validate optimal exercise and weight loss treatments for overweight and obese adults with knee osteoarthritis. Arthritis care & research, 2020. doi: https://doi.org/10.1002/acr.24179.

Michael I. Jordan. Artificial intelligence—the revolution hasn't happened yet. Harvard Data Science Review, 1(1), 7 2019. doi: 10.1162/99608f92.f06c6e61. URL https://hdsr.mitpress.mit.edu/pub/wot7mkc1. https://hdsr.mitpress.mit.edu/pub/wot7mkc1.

Anna R Kahkoska, Madison E Watts, Kimberly A Driscoll, Franziska K Bishop, Paul Mihas, Joan Thomas, Jennifer R Law, Nina Jain, and Elizabeth J Mayer-Davis. Understanding antagonism and synergism: A qualitative assessment of weight management in youth with type 1 diabetes mellitus. Obesity medicine, 9:21–31, 2018.

Michael R Kosorok and Eric B Laber. Precision medicine. Annual review of statistics and its application, 6:263–286, 2019.

Lenna L Liu, Jean M Lawrence, Cralen Davis, Angela D Liese, David J Pettitt, Catherine Pihoker, Dana Dabelea, Richard Hamman, Beth Waitzfelder, Henry S Kahn, et al. Prevalence of overweight and obesity in youth with diabetes in usa: the search for diabetes in youth study. Pediatric diabetes, 11(1):4–11, 2010.

Daniel J Luckett, Eric B Laber, Anna R Kahkoska, David M Maahs, Elizabeth Mayer-Davis, and Michael R Kosorok. Estimating dynamic treatment regimes in mobile health using v-learning. Journal of the American Statistical Association, 115(530), 2019.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, pages 4766 – 4775, 2017.

Donald Martin Jr, Vinod Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S Isaac. Participatory problem formulation for fairer machine learning through community based system dynamics. arXiv preprint arXiv:2005.07572, 2020a.

Donald Martin Jr, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S Isaac. Extending the machine learning abstraction boundary: A complex systems approach to incorporate societal context. arXiv preprint arXiv:2006.09663, 2020b.

Stephen P Messier, Shannon L Mihalko, Claudine Legault, Gary D Miller, Barbara J Nicklas, Paul DeVita, Daniel P Beavers, David J Hunter, Mary F Lyles, Felix Eckstein, et al. Effects of intensive diet and exercise on knee joint loads, inflammation, and clinical outcomes among overweight and obese adults with knee osteoarthritis: the idea randomized clinical trial. Jama, 310(12):1263–1273, 2013.

Susan A Murphy. Optimal dynamic treatment regimes. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(2):331–355, 2003.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464): 447–453, October 2019. doi: 10.1126/science.aax2342.

Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. Towards the science of security and privacy in machine learning. arXiv preprint arXiv:1611.03814, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pages 1135–1144, New York, NY, USA, August 2016. Association for Computing Machinery. doi: 10.1145/2939672.2939778.

Michael C Riddell, Ian W Gallen, Carmel E Smart, Craig E Taplin, Peter Adolfsson, Alistair N Lumb, Aaron Kowalski, Remi Rabasa-Lhoret, Rory J McCrimmon, Carin Hume, et al. Exercise management in type 1 diabetes: a consensus statement. The lancet Diabetes & endocrinology, 5(5):377–390, 2017.

Michael C Riddell, Dessi P Zaharieva, Michael Tansey, Eva Tsalikian, Gil Admon, Zoey Li, Craig Kollman, and Roy W Beck. Individual glucose responses to prolonged moderate intensity aerobic exercise in adolescents with type 1 diabetes: the higher they start, the harder they fall. Pediatric diabetes, 20(1):99–106, 2019.

Whitney R Robinson, Audrey Renson, and Ashley I Naimi. Teaching yourself about structural racism will improve your machine learning. Biostatistics, 21(2): 339–344, 11 2019. ISSN 1465-4644. doi: 10.1093/biostatistics/kxz040. URL https://doi.org/10.1093/biostatistics/kxz040.

Royal Statistical Society. Code of conduct, 2014. URL https://rss.org.uk/about/policy-and-guidelines/code-of-conduct/. Accessed: 2021-02-26.

Donald B Rubin. Bayesian inference for causal effects: The role of randomization. The Annals of statistics, pages 34–58, 1978.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association, 100(469):322–331, 2005.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5):206–215, 2019.

Cynthia Rudin and Berk Ustun. Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. Interfaces, 48(5):449–466, 2018.

Beth Schurman. The Framework for FDA's Real-World Evidence Program. Applied Clinical Trials, 2019. ISSN 10648542.

John Sperger, Nikki LB Freeman, Xiaotong Jiang, David Bang, Daniel de Marchi, and Michael R Kosorok. The future of precision health is data-driven decision support. Statistical Analysis and Data Mining: The ASA Data Science Journal, 13(6):537–543, 2020.

Dessi P Zaharieva, Ananta Addala, Kimber M Simmons, and David M Maahs. Weight management in youth with type 1 diabetes and obesity: Challenges and possible solutions. Current Obesity Reports, pages 1–12, 2020.

Tal Zarsky. The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. Science, Technology, & Human Values, 41(1):118–132, January 2016. doi: 10.1177/0162243915605575.

Yichi Zhang, Eric B Laber, Marie Davidian, and Anastasios A Tsiatis. Interpretable dynamic treatment regimes. Journal of the American Statistical Association, 113(524): 1541–1549, 2018.

Chunhui Zhao, Youxian Sun, and Luping Zhao. Interindividual glucose dynamics in different frequency bands for online prediction of subcutaneous glucose concentration in type 1 diabetic subjects. AIChE Journal, 59(11):4228–4240, 2013.