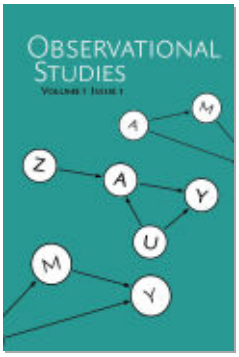
Comments on "Two Cultures": What have changed over 20 years?

Xuming He, Jingshen Wang



Observational Studies, Volume 7, Issue 1, 2021, pp. 99-106 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2021.0026*

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

Comments on "Two Cultures": What have changed over 20 years?

Xuming He **xmhe@umich.edu**
Department of Statistics
University of Michigan
273 West Hall, Ann Arbor, MI 48109, USA

Jingshen Wang **jingshenwang@berkeley.edu**
Division of Biostatistics
University of California, Berkeley
5408 Berkeley Way West, CA 94710, USA

Abstract

Twenty years ago Breiman (2001) called to our attention a significant cultural division in modeling and data analysis between the stochastic data models and the algorithmic models. Out of his deep concern that the statistical community was so deeply and "almost exclusively" committed to the former, Breiman warned that we were losing our abilities to solve many real world problems. Breiman was not the first, and certainly not the only statistician, to sound the alarm; we may refer to none other than John Tukey who wrote almost 60 years ago "data analysis is intrinsically an empirical science". However, the bluntness and timeliness of Breiman's article made it uniquely influential. It prepared us for the data science era and encouraged a new generation of statisticians to embrace a more broadly defined discipline.

Some might argue that "The cultural division between these two statistical learning frameworks has been growing at a steady pace in recent years", to quote Mukhopadhyay and Wang (2020). In this commentary, we focus on some of the positive changes over the past 20 years and offer an optimistic outlook for our profession.

Keywords: Algorithm; Data models; Machine learning, Prediction.

1. Diverse evaluation metrics are taking root in statistics

Classical statistical theory relies mostly on probability or stochastic models and places much emphasis on the accuracy of parameter estimation (e.g., bias, variance, mean squared error, coverage probability), and the validity and the power of hypothesis testing on certain parameters of interest. However, we also routinely use predictive/forecasting accuracy, classification accuracy, individual/subgroup treatment effects, receiver operating characteristic curves and so on in statistical assessments, depending on which is more relevant to the problem being addressed.

The importance of prediction is gaining recognition. We found the use of "prediction error" in 16 articles (including comments and book reviews) published in *Journal of the American Statistical Association* over a 3-year period of 1998-2000, and the number rose more than quintupled to 86 over the most recent 3-year period (2018-2020). To put this in

perspective, the same search engine showed an increase from 79 articles to 138 (just shy of doubling) over the same time period for "logistic regression".

Prediction is certainly not the only thing that matters. We do not think it is fair or even possible to quantify how much we do in statistics should be about prediction accuracy. Instead, we ask whether more needs to be done to promote the use of modern statistical and machine learning methods in real world applications. We share our experience in two areas of applications.

Sport concussion concerns many but accurate diagnosis and good understanding of recovery and of longer term impact remain challenging. Because concussion is typically a result of biomechanical forces transmitted to the cerebral tissues from impacts to the head or torso, clinical researchers have asked whether biomechanical analyses of impacts can help in developing a diagnostic model and improving protective equipment for the athletes. Earlier work using tree-based methods (Broglio et al., 2010) to analyze data collected from the Head Impact Telemetry System helped quantify importance of the biomecehnical measurements and the relevant thresholds for concussive injuries sustained during high school football. Over the years we also learned that accuracy of concussion diagnosis should not rely on those measurements alone. Furthermore, machine learning methods including the deep neutral network model was introduced to clinical researchers for predicting time-to-recovery in pediatric concussion patients; see Walker et al. (2019). Genetic Fuzzy Trees (GFT), a modern artificial intelligence tool, was employed in Fleck et al. (2020) to predict post-concussion symptom recovery in adolescents. When compared with common linear classification methods, the GFT was shown to have significantly better overall accuracy than most of its competitors based on cross-validated measures of sensitivity and specificity. In this particular study the overall accuracy based on Discriminant Function Analysis (DFA) and Random Forest was 54.4% and 57.1%, respectively, but increased to 62.3% when the GFT was used. Studies of this nature in concussion research are still limited, but our engagement in the Michigan Concussion Center has convinced us that our colleagues often turn to us as statisticians for help with both exploratory and confirmatory analysis, and few of them consider us as experts of "one culture". They talk to us about predictive accuracy as well as significance tests under interpretable models.

Cardiovascular disease (CVD) is a major cause of death worldwide. Multiple risk factors (i.e, age, gender, smoking status, body mass index, systolic blood pressure, history of diabetes, and reception of treatments for hypertension) are associated with CVD. Appropriately accounting for these risk factors is known to improve accuracy of CVD risk prediction and hence extend lives. Conventional methods including the Cox proportional hazard models are still being frequently employed to estimate the CVD risk based on these familiar risk factors (see Pylypchuk et al., 2018; Kaptoge et al., 2019, for example). In fact, many current guidelines often recommend calculating the CVD risk score based on a limited number of predictors: the 2010 American College of Cardiology/American Heart Association (ACC/AHA) guideline recommended the Framingham Risk Score (Greenland et al., 2010), the 2016 United Kingdom National Institute for Health and Care Excellence (NICE) guidelines used the QRISK3 score (Hippisley-Cox et al., 2017), and the 2016 European Society of Cardiology (ESC) guidelines included the SCORE model (Members et al., 2016).

Because of restrictive modelling assumptions adopted in traditional CVD risk score models, these approaches generally exhibit modest predictive performance. In recent years, modern machine learning tools have begun to be employed for predicting the CVD risk in the hope of agnostically discovering novel risk predictors and learning the complex interactions between them (see Krittanawong et al., 2020, for a comprehensive review). For example, "AutoPrognosis", an automated clinical prognostic modeling tool, was able to correctly predict 368 more cases out of 4,801 CVD cases than a traditional approach in analyzing UK Biobank data; see Alaa et al. (2019). It was shown in Sung et al. (2019) that a deep learning prediction model provided some promising results in predicting the CVD risk. There, the deep learning model had time-dependent "Area under the ROC Curve" (AUC) of as high as 0.94 for females and 0.96 for males, while the classical Cox regression model showed time-dependent AUC of up to 0.79 for females and 0.75 for males. In both studies cited above, the contribution of each variable to the prediction was used to rank the possible risk factors, but a variable that contributes more to prediction is not necessarily causal. Quantification of a risk factor via a causality study in the machine learning framework remains underdeveloped. Nevertheless, the capacity of machine learning algorithms to model highly interactive complex data structures in CVD has gained momentum. By Google Scholar counts, we found that around 10% of the publications about CVD discussed machine learning methods in 2010, and the share has risen to about one third by now.

2. Data models and algorithmic models are no longer easily distinguishable

Breiman made a convincing case that stochastic data models used in classical statistics are not the best, or even appropriate, models for many real world problems. Algorithmic models often offer a much more flexible framework for data analysis. We would like to emphasize that data models in statistics are often used as working models to generate a procedure or algorithm. They do not need to be correct or even accurate descriptions of the true nature, either for prediction or for inference. In this sense, data models and algorithmic models are often two starting points that could end up with friendly hand-shaking.

There are many examples where mis-specification of the data models are taken into account in inference. Generalized estimating equations (GEE), originally designed to account for a possible unknown correlation between outcomes (Zeger and Liang, 1986), relies on specification of only the moments in a data model. By using a robust variance-covariance estimates, statistical inference on the model parameters remain asymptotically valid without stringent assumptions assumed on part of the models. Another example is the posterior inference in a Bayesian framework. Yang et al. (2016) showed that by adopting a convenient working likelihood for quantile regression, Markov Chain Monte Carlo algorithms can then be used to draw from the posterior distribution to form a basis for asymptotically valid inference on the quantile regression parameter; here the assumed data model is really used as an input to the Bayesian computational framework, and the entire inferential process can be viewed as an algorithm.

Taking it a little further, we have mixture of experts (Jacobs et al., 1991) that starts with Gaussian mixture models. The Gaussian components may allow us to derive and

study certain properties of the model but we can take this supervised learning procedure as algorithmic.

In addition, current research in analyzing the CVD risk often finds data models and algorithmic models complimentary in identifying important risk factors. When predicting the risk of CVD by using deep learning models, Sung et al. (2019) ranked the influence of the input variables by using a Layer-wise Relevance Propagation (LRP), one of many explainable artificial intelligence techniques used in neural networks. They found that many risk factors (including age, gender, blood pressure, smoking, exercise) considered to be important in classical clinical studies were also shown to be important by deep learning models. Deep learning has its own limitation in providing the effect size of those risk factors due to hidden layers used in LRP. Nevertheless, Alaa et al. (2019) found that the algorithmic models helped uncover novel risk factors for CVDs that have not been considered in traditional prospective studies when analyzing the UK Biobank data. For women, AutoPrognosis identified "ankle spacing width" as an important risk factor. Clinically, this may be linked to symptoms of poor circulation, such as swollen legs, which could be predictive of future CVD events. For patients with history of diabetes, AutoPrognosis has not only maintained high predictive accuracy but also revealed diabetes-specific risk factors (such as microalbuminuria) that were not previously captured by the traditional data models.

We certainly do not intend to claim that the difference between the data models and the algorithmic models no longer exists. It remains obvious that both types of models can be mis-used. In fact, dangers have persisted in many statistical analyses that rely too much on blind trusts of the data models, which has led to the reproducibility and replicability crisis as some have called it. There are also many instances where blind use of algorithmic models and machine learning methods have failed us. The Google Flu Trends (Butler, 2013), IBM's Watson supercomputer cancer treatment recommendations (Ross and Swetlitz, 2018), and Apple Card Algorithm (Vigdor, 2019; Pena et al., 2020) are just some of well-publicized examples. We do not think that these problems argue against either type of models; rather their co-existence and further promotion of their better use make statistics a better playground. This leads us to the next point.

3. Statistical learning needs domain expertise

Data analysis is not just about data. Instead, any modeling attempt needs to incorporate domain knowledge and expertise wherever possible. We think this is a bigger issue than which modeling approach one takes.

Take statistical downscaling as an example. Statistical downscaling aims to localize global or regional climate model projections to assess the potential impact of climate changes based on data-driven models. It requires quantifying a relationship between climate model outputs (X) and local observations (Y) from the past. One would naturally attempt to regress Y on X to find a relationship that can be used for future projections of Y based on the climate model output. However, the projections from typical climate models are about "climate", not daily/weekly/monthly measurements that pair with the local observations in real time. Without the domain knowledge that we had to learn from the atmospheric scientists, we would run the risk of making poor choices of statistical modeling tools, whether it is a simple linear regression or a more complicated machine learning method. In fact, we

need to turn to regression-type modeling with asynchronous measurements as demonstrated in He et al. (2012) and Stoner et al. (2013).

In computational medicine, it is no surprise that machine learning algorithms without incorporating domain knowledge can make mistakes that would appear trivial to domain experts. In a recent study, CheXNet (Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning), Rajpurkar et al. (2017) observed that a convolutional neural network (CNN) outperformed radiologists in overall diagnostic accuracy. A subsequent study (Zech et al., 2018) revealed that the CNN built in CheXNet suffers from high generalization bias. Part of the reasons is that the CNN chooses some hospital site-specific variables over patient underlying pathology as predictive variables. Such a choice was made based on a strong correlation displayed in CheXNet between the pneumonia prevalence and those site-specific variables. In addition to the site-specific confounding variables that threaten the generalizability of the CNN, there can be other factors related to medical management that undermines the clinical applicability of statistical learning methods. Oakden-Rayner (2020) pointed out since chest tubes are frequently used to treat pneumothorax, a CNN without context-oriented modelling may not be able to diagnose patients with pneumothorax since they lacked a chest tube. To make statistical learning methods more trustworthy in medical research, adapting to domain knowledge is an important step towards building stronger and more robust models. Whether one calls this "interpretable machine learning" or "explainable AI", we need to be embedded in the relevant domain of science, business, or medicine.

Data-driven models and prediction have proven to be useful but not without peril. Things could go wrong easily if data are taken out of context. To conclude, we reiterate some of the points made earlier. First, data alone, no matter how big, may not tell the whole story. Any prediction trained on the existing data without sanity checks based on domain knowledge or even common sense could turn into a fiasco. Second, a good model needs to incorporate the dynamic nature of the world, where non-stationarity and heterogeneity of various forms cannot be ignored. Generalization errors that do not look beyond the past data are not always trusty measures. Third, we need to understand the limitations of what we can do with data in any approach to modeling. Context-oriented modeling is key to make statistical learning more powerful. Looking back, we statisticians have made significant progresses in the right direction, and we certainly expect more to come.

Acknowledgments

The research is partially supported by the National Science Foundation Awards DMS-1914496 and DMS-2015325.

References

Ahmed M Alaa, Thomas Bolton, Emanuele Di Angelantonio, James HF Rudd, and Mihaela van der Schaar. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK biobank participants. *PloS One*, 14(5):e0213653, 2019.

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

Steven P Broglio, Brock Schnebel, Jacob J Sosnoff, Sunghoon Shin, Xingdong Feng, Xuming He, and Jerrad Zimmerman. The biomechanical properties of concussions in high school football. *Medicine and Science in Sports and Exercise*, 42(11):2064, 2010.

Declan Butler. When Google got flu wrong. *Nature News*, 494(7436):155, 2013.

David E Fleck, Nicholas Ernest, Ruth Asch, Caleb M Adler, Kelly Cohen, Weihong Yuan, Brandon Kunkel, Robert Krikorian, Shari L Wade, and Lynn Babcock. Predicting post-concussion symptom recovery in adolescents using a novel artificial intelligence. *Journal of Neurotrauma*, 2020.

Philip Greenland, Joseph S Alpert, George A Beller, Emelia J Benjamin, Matthew J Budoff, Zahi A Fayad, Elyse Foster, Mark A Hlatky, John McB Hodgson, and Frederick G Kushner. 2010 ACCF/AHA guideline for assessment of cardiovascular risk in asymptomatic adults: a report of the american college of cardiology foundation/american heart association task force on practice guidelines. *Circulation*, 122(25):e584–e636, 2010.

Xuming He, Yunwen Yang, and Jingfei Zhang. Bivariate downscaling with asynchronous measurements. *Journal of Agricultural, Biological, and Environmental Statistics*, 17(3): 476–489, 2012.

Julia Hippisley-Cox, Carol Coupland, and Peter Brindle. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*, 357, 2017.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Stephen Kaptoge, Lisa Pennells, Dirk De Bacquer, Marie Therese Cooney, Maryam Kavousi, Gretchen Stevens, Leanne Margaret Riley, Stefan Savin, Taskeen Khan, and Servet Altay. World health organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. *The Lancet Global Health*, 7(10):e1332–e1345, 2019.

Chayakrit Krittanawong, Hafeez Ul Hassan Virk, Sripal Bangalore, Zhen Wang, Kipp W Johnson, Rachel Pinotti, HongJu Zhang, Scott Kaplin, Bharat Narasimhan, and Takeshi Kitai. Machine learning prediction in cardiovascular diseases: a meta-analysis. *Scientific Reports*, 10(1):1–11, 2020.

Task Force Members, Massimo F Piepoli, Arno W Hoes, Stefan Agewall, Christian Albus, Carlos Brotons, Alberico L Catapano, Marie-Therese Cooney, Ugo Corrà, and Cosyns.

2016 european guidelines on cardiovascular disease prevention in clinical practice. *European Journal of Preventive Cardiology*, 23(11):NP1–NP96, 2016.

Subhadeep Mukhopadhyay and Kaijun Wang. Breiman's "two cultures" revisited and reconciled. *arXiv preprint arXiv:2005.13596*, 2020.

Luke Oakden-Rayner. Exploring large-scale public medical image datasets. *Academic Radiology*, 27(1):106–112, 2020.

Alejandro Pena, Ignacio Serna, Aythami Morales, and Julian Fierrez. Bias in multimodal AI: testbed for fair automatic recruitment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–29, 2020.

Romana Pylypchuk, Sue Wells, Andrew Kerr, Katrina Poppe, Tania Riddell, Matire Harwood, Dan Exeter, Suneela Mehta, Corina Grey, and Billy P Wu. Cardiovascular disease risk prediction equations in 400 000 primary care patients in new zealand: a derivation and validation study. *The Lancet*, 391(10133):1897–1907, 2018.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, and Katie Shpanskaya. Chexnet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Casey Ross and Ike Swetlitz. IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. *Stat*, 25, 2018.

Anne MK Stoner, Katharine Hayhoe, Xiaohui Yang, and Donald J Wuebbles. An asynchronous regional regression model for statistical downscaling of daily climate variables. *International Journal of Climatology*, 33(11):2473–2494, 2013.

Ji Min Sung, In-Jeong Cho, David Sung, Sunhee Kim, Hyeon Chang Kim, Myeong-Hun Chae, Maryam Kavousi, Oscar L Rueda-Ochoa, M Arfan Ikram, and Oscar H Franco. Development and verification of prediction models for preventing cardiovascular diseases. *PloS One*, 14(9):e0222809, 2019.

Neil Vigdor. Apple card investigated after gender discrimination complaints. *The New York Times*, 10, 2019.

Gregory Walker, Adam Kiefer, Nathaniel Richards, Emma Klug, Christy Reed, Paula Silva, and Kelly Cohen. Machine learning for concussion recovery prognosis: a novel tool to empower proactive physician treatments, 2019.

Yunwen Yang, Huixia Judy Wang, and Xuming He. Posterior inference in bayesian quantile regression with asymmetric laplace likelihood. *International Statistical Review*, 84(3): 327–344, 2016.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Medicine*, 15(11): e1002683, 2018.

Scott L Zeger and Kung-Yee Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, pages 121–130, 1986.