Chapter 5

Preventing Disparities: Bayesian and Frequentist Methods for Assessing Fairness in Machine-Learning Decision-Support Models

Douglas S. McNair

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.73176

Abstract

Machine-learning (ML) methods are finding increasing application to guide human decision-making in many fields. Such guidance can have important consequences, including treatments and outcomes in health care. Recently, growing attention has focused on the potential that machine-learning might automatically learn unjust or discriminatory, but unrecognized or undisclosed, patterns that are manifested in available observational data and the human processes that gave rise to them, and thereby inadvertently perpetuating and propagating injustices that are embodied in the historical data. We applied two frequentist methods that have long been utilized in the courts and elsewhere for the purpose of ascertaining fairness (Cochran-Mantel-Haenszel test and beta regression) and one Bayesian method (Bayesian Model Averaging). These methods revealed that our ML model for guiding physicians' prescribing discharge beta-blocker medication for post-coronary artery bypass patients do not manifest significant untoward race-associated disparity. The methods also showed that our ML model for directing repeat performance of MRI imaging in children with medulloblastoma did manifest racial disparities that are likely associated with ethnic differences in informed consent and desire for information in the context of serious malignancies. The relevance of these methods to ascertaining and assuring fairness in other ML-based decision-support model-development and -curation contexts is discussed.

Keywords: fairness, machine-learning, Bayesian model averaging, bias, variables selection

1. Introduction

With regard to cognitive computing and machine-learning (ML)-based decision-support tools, there is an emerging need for ethical reasoning about Big Data beyond privacy [1–3].

Recent definitions of 'algorithmic fairness' [4–7] assert that similar individuals should be treated similarly. Such metrics comport with conventional lay-persons' sense of the meaning of fairness. Algorithmic fairness definitions presuppose the existence of a use-case-specific metric on individuals and propose that fair algorithms should satisfy a Lipschitz condition with respect to this metric. However, such definitions for algorithms and artificial intelligence tools have not yet been aligned with existing statistical methods that have been established in the legal and regulatory communities. Furthermore, no generally accepted standards yet exist for ascertaining the presence or absence of disparities in machine-learning (ML) models that have been learned from historical observational data. There is a serious concern among policy-makers and members of the public that the rapid growth of ML may lead to the systematic promulgation of "bad models" that inculcate past injustices in subsequent decision-making going forward [8–31].

Such concerns are heightened in the context of life-critical medical and surgical treatment. In one illustrative medical example, beta-blocker medications have been found to be important in the treatment of myocardial infarction and in coronary artery bypass (CAB) surgery in that they have been shown to decrease mortality. Their benefit is derived not only from improving the myocardial oxygen supply-demand balance but also from their ability to inhibit subsequent cardiac ventricular remodeling, mitigation of platelet activation, decrease in peripheral vascular resistance (PVR) and decrease in hemodynamic stress on the arterial wall, increase in diastolic coronary artery flow, membrane stabilization and shortening of the heart rate-corrected QT interval (QT_c) [32], prevention of atrial fibrillation and other arrhythmias, and other mechanisms.

In another typical example in clinical medicine, serial repeated MRI scans of the head and spinal cord have been found to be relevant in the ongoing management of medulloblastoma [33]. As with any cancer, early detection and ongoing follow-up monitoring are essential to achieving a positive outcome. With its multi-planar capability and excellent high spatial resolution, MRI is the preferred imaging modality in the follow-up to assess response to treatment. The efficacy of repeated MRI scans is presently uncertain as regards improvement of survival or other outcomes. However, there can be considerable psychological value that attaches to finding that a repeat scan is negative for recurrence, progression, or metastasis of the cancer, and repeat scans are routinely performed at regular intervals on this empirical basis, motivated by the wish to provide knowledge and reassurance. Conversely, the MRI-informed discovery of recurrence, progression, or metastasis is a much-feared possibility for parents of children with medulloblastoma, insofar as this finding portends shortened life-expectancy for the child and diminution of hope. In certain contexts, then, there is a disinclination to perform exams that could lead to bad news, for which there may be no effective mitigations or treatment options.

2. Background and methodology

Avoiding Type II (false-negative) errors is paramount in machine-learning model quality assurance and fairness determinations. Following this spirit, we have recently developed a

new framework for statistically ascertaining ML model fairness. The purpose of this chapter is to introduce the new three-method framework to the machine-learning community and illustrate its use with two practical examples from clinical medicine specialties (namely cardiology and oncology). Our method involves joint application of the following methods to ML model-training and -test data, where the test data may be either (a) data arising from natural decision-making unaided by the ML model or (b) data arising from decision-making where human users are assisted by the ML model:

1. The Cochran-Mantel-Haenszel test;
2. Beta regression; and
3. Bayesian model averaging (BMA).

If the p-values for all three methods are non-significant, then the ML model is declared to be provisionally fair. However, if any of these methods show that statistically significant ($p < 0.05$) bias exists depending on one or more stratification variables, then the model is declared to have failed fairness checking, the model is placed into a "hold" status and not released, and further investigation is initiated into the nature of the detected bias and its possible causes.

To date, a majority of the more than 50 predictive mathematical models that have been developed and deployed by the author's team are ML models. The discovery, development, and validation of the models have primarily been performed using a HIPAA-compliant, de-identified and PHI-free, epsilon differential privacy-protected, secondary-use-assented, EHR-derived, ontology-mapped, longitudinally electronic master person identifier (eMPI) linked repository of the serial care-episode health records associated with 100% of patients cared for at 814 U.S. health institutions who have established HIPAA business-associate agreements and data-rights agreements with our corporation. This data warehouse currently comprises more than 153 million distinct persons' longitudinal records and more than 400 million episodes of care from January 1, 2000 to the present time. New case material accrues into the data warehouse from each of the contributing health networks' and institutions' systems on a daily basis, encrypted end-to-end, and auto-mapped to a standard ontology and pre-cleaned upon arrival. The data warehouse is not a "claims" dataset but instead encompasses a majority of the content of the patients' EHR records, from flowsheet and monitoring data and waveforms, to all medications dispenses and prescriptions, all lab results, all procedures, all problem list entries and diagnoses, and all claims—with each data element or item or transaction date-timestamped with minute-level time precision and with successive episodes for a given person longitudinally linked via a key that is encrypted from the eMPI. A typical ML project for us begins with a cohort extracted from the data warehouse. Cohorts for studies we undertake tend to comprise from 20,000 to several million cases and a comparable number of controls, all meeting inclusion-exclusion criteria for the project and governed by a project specification and written, version-controlled protocol. The datasets comprised of these cohorts of historical, outcomes-labeled, de-identified cases and controls are separated into randomized, independent "training" and "test"

subsets. A typical ML project for us begins with several hundred input data variables or document types selected from the EHR data model, which includes more than 10,000 data type categories.

2.1. Cohort selection

Two representative examples serve to illustrate the application of Bayesian and frequentist methods for assessing fairness in ML models, one involving a very large cohort (beta-blocker usage in coronary artery disease post-coronary artery bypass (CAB)) and one involving a comparatively small cohort (MRI in pediatric medulloblastoma (brain cancer)).

A post-CAB cohort included those cases who were discharged alive with hospital LOS between 3 and 28 days, black or white race only, between January 1, 2012 and December 31, 2016, aged between 40 and 69 years at the time of CAB surgery, with no known prior use of beta-blocker within 1 year prior to CAB. Excluded were patients receiving percutaneous and MIDCAB (usage rates for which might be, or are, confounded by geography, operative risk and preoperative comorbidities, and other factors); in-hospital percutaneous coronary intervention (PCI), PCI to CAB conversion, urgent-emergent CAB; known prior AMI, prior PCI or prior CAB; patients with heart rate <45 bpm or AV block (ICD-10-CM diagnosis codes I44.x, I45.x; ICD-9-CM diagnosis codes 426.x); patients with implanted pacemaker; patients having eGFR <50 mL/min/1.73m^2; persons with previously diagnosed heart failure, asthma, or active malignancy; patients who were transferred to other medical institutions without discharge prescription; and patients at institutions having fewer than 100 open CAB cases annually meeting the criteria above during 2012–2016. Patients treated at a total of 14 out of 814 institutions participating in this data warehouse met the criteria for inclusion in the ML model development and analysis.

A medulloblastoma cohort included cases who were discharged alive, black or white race only, between January 1, 2000 and December 31, 2016, aged between 0 and 21 at the time of resection of the brain tumor. Patients treated at a total of 33 out of 814 institutions participating in this data warehouse met the criteria for inclusion in the ML model development and analysis.

2.2. Data extraction

Exploratory analyses to characterize available data often require full table scans, which, in conventional RDBMS tables having billions of rows, may entail runtimes of many hours, even with bitmapped indexes and careful query optimization. Laboratory tests and vital signs and flowsheet items in our data warehouse are each multi-billion-row tables. Premature dimensionality or cardinality reduction may interfere with discovering the best ML model. Therefore, a 64-node Hewlett Packard Vertica® system was the means whereby the data warehouse was physically stored for the present work. Extracts were performed using standard SQL queries on this massively parallel vertical database. Although many racial and ethnic categories were represented in the data warehouse, for the present work racial categories were restricted to black and white, for reasons of adequacy of sample size.

A total of 30,116 complete post-CAB cases were retained, and no imputation was used. Median age was 64 years and 13.6% were black, with M:F ratio 2.57. From this extract, males were retained for analysis (median age 64 years, 11.1% black). Matching was performed on a per-hospital basis by race in a 1:9 ratio (Black:White), to minimize bias arising from regional differences in the prevalence of Black individuals. Matching was performed on U.S. census division (nine geographic regions) and on age with 5-year binning. Matching was additionally performed on diabetic status. This resulted in of 11,358 actual cases used for subsequent training dataset modeling and analysis. The remainder of the data was used as an independent test dataset.

A total of 1207 medulloblastoma cases were retained. Median age was 5.8 and 15.2% were black, with M:F ratio 1.71.

2.3. Feature selection

In our two examples, exploratory machine-learning, including logistic regression, was performed using raw data comprised of 326 data elements from the de-identified EHR-derived extracts, supplemented by derived variables that were transformed. The LASSO procedure [34] was used for dimensionality reduction. Predictor variables with a category-wise Wald test p-value ≤0.05 were retained in the models.

In the post-CAB beta-blocker example, transformed continuous-variable features (6) in the model included: ln(inter-beat interval), RMSSD(HR), ln(nbr_dx), ln(nbr_meds), ln(LOS_days), and ln(AST). Transformed binomial-variable features included the following: AST/ALT <1.1, max(HR) < 110, range(HR 48 hr prior to discharge) > 30 bpm, range(RR 48 hr prior to discharge) > 18 bpm, range(MAP 48 hr prior to discharge) > 22 mmHg, max(SBP) during hospital stay) < 150 mmHg; diabetes; concomitant calcium channel blocker; perioperative inotrope or mechanical circulatory assist; concomitant CYP2D6 substrate or inhibitor (esp. antidepressants, antipsychotics, COX-2 inhibitors, amiodarone, or cimetidine); history of substance abuse; and history of syncope, vertigo, postural hypotension, or falling.

In the medulloblastoma repeat MRI example, binomial-variable features included the following: clinical trial enrollment, prior evidence of recurrence or metastasis of tumor, renal impairment such as would be a safety contraindication for MRI contrast, high-risk histology, SHH or WNT genomics, tumor extent at resection, PFS duration, recent ^{99m}Tc scan, recent ^{123}I-mIGB scan, and public payor (Medicaid).

3. Comparing model-guided and natural decision-making

Personalized patient care decisions require considering numerous clinical information items and weighing and combining them according to patient-specific risks and likely benefits. Additionally, considerations of disease etiology and progression as well as on comorbid conditions and concomitant medications or prior treatments that may affect the underlying biological processes or constrain subsequent therapeutic options are required. Yet further,

guidelines regarding treatment modalities, risk factors, complications, patient caregiver support, living situation, and costs also influence care decisions. Natural, model-unassisted decision-making yields therapeutic treatment allocations that are the basis of the initial ML models. However, once one or more ML models are deployed and integrated with the users' workflow and decision-making, the guidance and evidence that the models present to the users tends to alter their decision-making and change the rates of allocating specific treatments or diagnostic procedures to individual patients. It is important to assess the fairness of ML models not only prior to their initial commissioning and deployment but also to reassess model fairness in a periodic and ongoing manner post-deployment. Depending on the degree to which an ML model influences users' decision-making it is possible that differences between strata may increase during deployment, and the model-guided data that accrues during the post-deployment period may cause later versions of the ML model to manifest statistically significant biases that were not present in the initial ML model version that was based on purely natural decisional data.

In the post-CAB beta-blocker example, the ML score output would later be consumed by prescribers in computerized physician order-entry (CPOE) apps used to advise the implementing of care in the perioperative CAB patients. Markov Chain Monte Carlo sampling of 11,358 cases in the "training" dataset was performed to determine the rate of historical discharge beta-blocker usage in each decile of ML-model-generated score values. In the serial MRI medulloblastoma follow-up example, the ML score output would later be consumed by prescribers in computerized physician order-entry (CPOE) apps used to advise the implementing of care in pediatric medulloblastoma patients. Markov Chain Monte Carlo sampling of 1207 cases in the "training" dataset was performed to determine the rate of historical serial MRI usage in each decile of ML-model-generated score values.

4. Evaluation approach

The purpose of fairness auditing in our two examples was to examine the questions (1) whether black patients were less likely to receive beneficial therapy or diagnostic procedures when compared with white patients and (2) whether, in connection with ML model-training on observational data from a large, representative collection of hospitals, an ML decision-support model would manifest a statistically significant untoward disparity of therapy or diagnostic procedures prescribing based on race. It was first necessary to determine whether the ML models were adequately calibrated in 'test' cohorts different from the ML model-discovery 'training' cohorts. Controlling for age distribution, geographic differences, gender, common contraindications for the treatment-of-interest (discharge beta-blocker post-CAB), and other factors [34–42] is important, to insure adequate statistical power for these assessments and to mitigate confounding [27, 43]. Establishing that the ML model was adequately well-calibrated for each racial group prior to performing procedures to evaluate the presence of discrimination or disparities was performed using the Hosmer-Lemeshow test by model score deciles. For black subjects, the model's HL was $\chi^2 = 10.9$, df = 8, p-value = 0.21, while for white subjects, HL $\chi^2 = 10.1$, and p-value = 0.26, confirming that the ML model scores showed

good calibration across the deciles of score values providing the recommendations for discharge beta-blocker prescribing. The distribution of discharge beta-blocker medications in the subset of the cohort who received them was as follows: metoprolol, 68.2%; carvedilol, 14.1%; labetalol, 11.5%; atenolol, 4.7%; propranolol, 0.87%; nebivolol, 0.28%; bisoprolol, 0.17%; nadolol, 0.08%; acebutolol, 0.04%; and pindolol, 0.02%. This distribution is consistent with recently published guidelines [44–47]. Kruskal-Wallis non-parametric ANOVA revealed no statistically significant racial group-associated differences in the proportions of these categories of beta-blockers. Corresponding controlling for age distribution and other factors was performed for the medulloblastoma example. Hosmer-Lemeshow evidence of model calibration was confirmed for the medulloblastoma ML model. With calibration determined to be adequate, we then proceeded to evaluate potential ML model biases using three methods: Cochran-Mantel-Haenszel test; beta regression; and Bayesian Model Averaging.

4.1. Cochran-Mantel-Haenszel test

Linear regression with normally distributed errors is probably the most commonly used analysis tool in applied statistics. The pervasiveness of linear regression is based on the fact that random variations in observed data can frequently be well-approximated by a normal distribution with constant variance. If the response variable in a regression model is a rate or percentage, however, the assumption of normally distributed errors is not valid. Because the analysis of rates and proportions is an important issue for many applications, establishing statistically valid analysis tools for dependent variables whose values are on the bounded interval [0,1) has high importance. This is particularly so in applications that assesses the fairness and equitability of proportions of allocated services or resources, including allocations that are mediated by decision-support tools and artificial intelligence (AI) models originating in ML from existing data. Such models aim to represent the relationship between a binary exposure (exposed vs. unexposed) and a binary outcome (success vs. failure). Sometimes the relationship between the two binary variables is influenced by another variable (or variables). One way to adjust for such influence is to stratify on that variable and perform stratified analysis.

The Cochran-Mantel-Haenszel test (CMH) is a test of the similarity of the mean rank (across the outcome scale) for groups in stratified 2 × 2 tables with possibly unbalanced stratum sizes and unbalanced group sizes within each stratum. The CMH test has the advantage of only moderate assumptions for calculating the p-value, namely, that the conditional odds ratios of the strata are in the same direction and similar in magnitude.

Cochran-Mantel-Haenszel (CMH) procedure tests the homogeneity of population proportions after taking into account other factors. The CMH test has been utilized for many years in the courts and by regulatory agencies [48–52]. The "training" and "test" data were arranged as a 2 × 2 × N arrays, where race and beta-blocker status comprised the first two dimensions and hospital was the third dimension. In this manner CMH examines one factor (race) and one outcome (discharge beta-blocker), across N subgroups (hospitals). The CMH chi-square tests if there is an interaction or association between the 2 × 2 rows and columns across the N categories. The null hypothesis is that the pooled odds ratio is equal to 1.0, there is no interaction

between rows and columns. Rejection of H_0 indicates that interaction exists. Calculation of the CMH test may be performed via the cmh.test() function in the R package 'lawstat' (https://cran.r-project.org/package=lawstat) or by other conventional means.

In the post-CAB beta-blocker analysis (**Table 1**), the CMH statistic = 5.84, df = 1, p-value = 0.016, MH Estimate = 1.23, Pooled Odds Ratio = 1.35, such that, rather than representing a disadvantage, black race in this male cohort conferred a slight advantage, with a modest increase in the likelihood of receiving discharge beta-blocker post-CAB compared to men who were white.

In the medulloblastoma repeat MRI analysis (**Table 2**), the CMH statistic = 39.8, df = 1, p-value <0.0001, MH estimate = 0.33, Pooled odds ratio = 0.35, such that children of black race in this cohort have a statistically lower likelihood of receiving serial MRI exams compared to children who were white.

4.2. Beta regression

Note that if the we see very different odds ratios for the strata, that suggest the variable used to separate the data into strata (race, in these examples) is a confounder and, if so, the Mantel-Haenszel odds ratio is not a valid measure of significance. To test whether the odds ratios in the different strata are different, we calculate Tarone's test of homogeneity using the rma.mh() function from the R package metafor. If some odds ratios are <1 and other odds ratios are >1, or if the Tarone test p-value <0.05, then the CMH test is not valid or appropriate. Thus, a disadvantage of CMH is that the circumstance of violation of its assumptions does occur comparatively often (for example, if the stratifying factor can confer protection for one value and excess risk for another value). Therefore, we sought additional methods that do not have this limitation.

One such alternative method that is able to address model rates and proportions is beta regression. Beta regression is based on the assumption that the response is beta-distributed on the unit interval [0,1). The beta density can assume a number of different shapes depending on the combination of parameter values, including left- and right-skewed or the flat shape of the uniform density. Beta regression models can allow for heteroskedasticity and can accommodate both variable dispersion and asymmetrical distributions. An additional advantage is that the regression parameters are interpretable in terms of the mean of the outcome variable.

The measure of association between the predictor variables and the outcome from the beta regression is expressed as a relative proportion ratio [53–56]. Beta regression is a model of the mean of the dependent variable y (likelihood of discharge beta-blocker) conditioned on covariates x (race, ML model-guided recommendation for beta-blocker, and the interaction

Race	Beta-blocker +	Beta-blocker -	Prevalence (BB+)	
			Actual (Training) (%)	Model-guided (Test) (%)
Black	953	307	75.6	73.3
White	7039	3059	69.7	67.8

Table 1. Prevalence of discharge beta-blocker utilization, post-CAB.

Race	MRI +	MRI -	Prevalence (MRI+)	
			Actual (Training) (%)	Model-guided (Test) (%)
Black	112	71	61.3	63.4
White	837	187	81.7	80.9

Table 2. Prevalence of serial MRI utilization, post-medulloblastoma resection.

between these), which we denote by μ_x. Because y is on the open interval (0, 1), we must ensure that μ_x is also in [0, 1). We do this by using a link function for the conditional mean, denoted $g(\cdot)$. This is necessary because linear combinations of the covariates are not otherwise restricted to [0, 1). Beta regression is widely used because of its flexibility for modeling variables whose values are constrained to lie between 0 and 1 and because its predictions are confined to the same range [53, 54]. Beta regression models were proposed by Ferrari and Cribari-Neto [55, 56] and extended by Smithson and Verkuilen [54] to allow the scale parameter to depend on covariates. We have:

$$g(\mu_x) = x\beta, \text{ or, equivalently, } \mu_x = g^{-1}(x\beta) \tag{1}$$

where $g^{-1}(\cdot)$ is the inverse function of $g(\cdot)$. Here the default logit link implies that

$$ln\left\{\mu_x/(1-\mu_x)\right\} = x\beta, \text{ and that } \mu_x = exp(x\beta)/\left\{1+exp(x\beta)\right\}. \tag{2}$$

Using a link function to keep the conditional-mean model inside an interval is common in the statistical literature. The conditional variance of the beta distribution is:

$$var(y \mid x) = \left\{ \mu_x (1-\mu_x) \right\}/(1+\psi). \tag{3}$$

The parameter ψ is known as the scale factor because it rescales the conditional variance. We use the scale link to ensure that $\psi > 0$.

Beta regression models have applications in a variety of disciplines, such as economics, the social sciences, and health science. For example, in political science and in the law, beta regression has been utilized in determining noncompliance with antidiscrimination laws [52]. In psychology, Smithson [57] used beta regression to evaluate jurors' assessments of the probability of a defendant's guilt and their verdicts in trial courts. Beta regression has also been used to model quality-adjusted life years in health cost-effectiveness studies [58, 59].

Where necessary, outcome observations (the proportion of cases receiving discharge beta-blocker post-CAB) were transformed to the open unit interval (0, 1), adding a very small amount (0.001) to the zero-valued observations and subtracting the same amount from the one-valued observations. Beta regression was performed via the betareg() function in the R package 'betareg' (https://cran.r-project.org/package=betareg) but may also be accomplished by other similar algorithms in other statistics packages. Beta regression (**Table 3**) produces

Covariate	Estimate	Std error	p-Value
(Intercept)	−1.136	0.067	< 0.0001
Black	0.201	0.067	0.0026
Score_percentile	4.875	0.119	< 0.0001
Black:Score_percentile	−0.033#	0.118	0.7780
(phi)_(Intercept)	2.434	0.160	< 0.0001
(phi)_Black	−0.093	0.099	0.3466
(phi)_Score_percentile	2.394	0.255	< 0.0001

This shows that discharge beta-blocker rate increases with Score_percentile and is slightly higher (for blacks), and there is no significant interaction (annotated as #) between Score_percentile and black race. This evidence corroborates that from the Cochran-Mantel-Haenszel test, regarding the absence of disadvantage under the ML model for black men compared to white men for post-CAB discharge beta-blocker recommendation. Precision is asymmetric and heteroskedastic. Precision (phi) increases with Score_percentile.

Table 3. Beta regression of discharge beta-blocker utilization, post-CAB.

estimated coefficients of the covariates and an estimated scale parameter. The coefficient of the factor variable for race = Black is significant at the $p < 0.05$ level and positive. Thus we conclude that, rather than posing a hazard, Black race in these 14 institutions during this 5-year period, actually conferred a slight advantage in terms of the likelihood of a male patient's receiving standard-of-care discharge on a beta-blocker, status-post open coronary artery bypass.

Corresponding beta regression (**Table 4**) was performed for predictive recommendations from our second example ML model derived from repeat MRI pediatric medulloblastoma data from 33 institutions.

Covariate	Estimate	Std error	p-Value
(Intercept)	−2.399	0.220	< 0.0001
Black	0.036	0.219	0.7813
Score_percentile	5.030	0.293	< 0.0001
Black:Score_percentile	−0.387#	0.290	0.1821
(phi)_(Intercept)	2.232	0.289	< 0.0001
(phi)_Black	0.061	0.122	0.6162
(phi)_Score_percentile	−0.023	0.410	0.9562

This shows that the rate of serial MRI exams increases with Score_percentile and in the mean equation there is potentially a weak, mildly negative interaction (annotated as #) between Score_percentile and black race. This is weak evidence consistent with the hypothesis that a disparity may exist under our initial, empirically discovered ML model, between black and white children with medulloblastoma with regard to recommendation of serial MRI scans in treatment follow-up. Precision (phi) is not significantly asymmetric or heteroskedastic in this example dataset.

Table 4. Beta regression of serial MRI utilization, post-medulloblastoma resection.

4.3. Bayesian model averaging

In our experience, beta regression and CMH are sufficient for ascertaining the fairness of ML-derived models in many situations. However, if the strata are markedly unbalanced or if the data are not satisfactorily fitted by a beta distribution, these methods may give either false-positive or false-negative results. Also, percentage outcomes that are based on the binomial model are often overdispersed, meaning that they show a larger variability than expected by the binomial distribution. Beta regression models usually account for overdispersion by including the precision parameter phi to adjust the conditional variance of the percentage outcome, but this fixed parameterization involves an *ad hoc* choice by the analyst and may be unstable or yield poor goodness-of-fit when the data are heteroskedastic. Yet further, beta regression tends to require relatively large sample sizes to power interpretations of statistical significance. Therefore, we seek additional methods that are robust against these conditions. In that regard, Bayesian model averaging (BMA) offers particular advantages.

BMA is a relatively recently developed method that addresses model uncertainty in the canonical regression variables selection problem [60–64]. If we assume a linear model structure, where y is the dependent variable to be predicted, α_i are constants, β_i are coefficients, and ε is a normal IID error term with variance σ^2 then we have:

$$y = \alpha_i + X_i \beta_i + \epsilon, \quad \varepsilon \sim N(0, \sigma^2 I) \tag{4}$$

High dimensionality interferes with stable variables selection. Small cohort size or collinearity of potential explanatory variables in matrix X may increase the risk of over-fitting and retention of some variables $Xi \in \{X\}$ which should not be included in the model. Stepwise variables elimination starting from the null linear model that includes all variables may be statistically unsupportable if the cohort size is small.

BMA addresses the problem by estimating models for all, or a very large number of, possible combinations of $\{X\}$ and constructing a weighted average over all of them. If there are K potential variables, this means estimating 2^K variable combinations and therefore 2^K models. The model weights for model averaging arise from posterior model probabilities which, in turn, are denoted by Bayes' theorem:

$$p(M_i \mid y, X) = \frac{p(y \mid M_i, X)\, p(M_i)}{p(y \mid X)} = \frac{p(y \mid M_i, X)\, p(M_i)}{\sum_{j=1}^{2^K} p(y \mid M_j, X)\, p(M_j)} \tag{5}$$

Here, $p(y \mid X)$ is the integrated likelihood, which is constant over all models. Therefore, the posterior model probability (PMP) $p(M_i \mid y, X)$ is proportional to the marginal likelihood of the model $p(y \mid M_i, X)$ (the likelihood of the observed data, given model M_i) times a prior model probability $p(M_i)$; that is, how probable the machine-learning analyst believes model M_i to

be before looking at the data. Renormalization then leads to the PMPs and thus the model weighted posterior distribution for any statistic θ (for example, the coefficients β_i):

$$p(\theta \mid y, X) = \sum_{i=1}^{2^K} p(\theta \mid M_i, y, X)\, p(M_i \mid X, y) \tag{6}$$

The model prior $p(M_i)$ is elicited by the machine-learning researcher and reflects prior beliefs [65, 66], some of which may come from published research literature [67]. In the absence of other guidance or historical knowledge, a routine option is to assume a uniform prior probability for all models $p(M_i) \propto 1$ to represent the absence of a well-established prior.

The expressions for posterior distributions $p(\theta \mid M_i, y, X)$ and for marginal likelihoods $p(M_i \mid y, X)$ depend on the model estimation framework. Routine practice is to use a Bayesian linear model with a prior structure called Zellner's g prior [68, 69]. For each candidate model M_i a normal-distributed error structure is assumed, as in Eq. (1). The need to determine posterior distributions requires that one specify priors on the model parameters. In practice, one sets provisional priors on the constants and on error variance, typically distributed as $p(\alpha_i) \propto 1$, meaning complete prior uncertainty about the prior mean, and $p(\sigma) \propto \sigma^{-1}$.

The most influential prior is the one on the coefficients β_i. Before analyzing the data (y, X), the analyst proposes priors on the coefficients β_i, typically normally distributed with a specified mean and variance. In the context of ML model fairness evaluations, we assume a prior mean of zero for the coefficients to assert that not much is known about them. In our work, their variance structure is defined by Zellner's g:

$$\beta_i \mid g \sim N\left(0, \sigma^2 \left(\frac{1}{g} X_i' X_i\right)^{-1}\right) \tag{7}$$

The hyperparameter g embodies how certain the analyst is that coefficients are zero: A small g means small prior coefficient variance and therefore implies the analyst is quite certain that the coefficients are indeed approximately zero. By contrast, a large g means that the analyst is very uncertain about whether the variables' coefficients are statistically significant, as in the case of our work on ML model fairness evaluations with regard to racial bias.

In general, the more complicated the distribution of marginal likelihoods, the more difficulties a Bayesian (Gibbs, Markov Chain Monte Carlo) sampler will encounter before converging to a good approximation of posterior model probabilities (PMPs). The quality of approximation may be inferred from the number of times a model got drawn versus their actual marginal likelihoods. Partly for this reason, BMA retains a pre-specified number of models with the highest PMPs encountered during MCMC sampling, for which PMPs and draw counts are stored. Their respective distributions and their correlation indicate how well the sampler has converged. While BMA should usually compare as many models as possible, some considerations might dictate the restriction to a subspace of the 2^K models. By far the most common setting is to keep some regressors fixed in the model setting, and apply Bayesian Model uncertainty only to a subset of regressors. However, due to physical RAM memory limits, the sampling chain can retain fewer than 1000,000 of these models. Instead, BMA computes aggregate statistics on-the-fly, usually using iteration counts as surrogate model weights. For model convergence and some posterior statistics BMA retains only the 'top' (highest PMP)

models it encounters during the iterations executed. Since the time for updating the iteration counts for the 'top' models grows linearly with their number, the sampler becomes considerably slower the more 'top' models that are retained. Still, if they are sufficiently numerous, those best models can accurately represent most of posterior model cumulative probability. In this case, it is defensible to base posterior statistics on analytical likelihoods instead of MCMC frequencies.

For the post-CAB beta-blocker at discharge example, **Table 5** shows features of the 10 top-ranked models generated by BMA MCMC sampling, together with the cumulative inclusion probability for each feature summed over the models evaluated.

With regard to prescribing of beta-blocker medication at discharge from hospital post-CAB coronary revascularization, Bayesian model averaging yields evidence that models omitting race have higher Posterior Model Probability (PMP) than models that retain race as a feature, and race exhibits low inclusion probability. These findings are compatible with the results of CMH and beta regression and support the hypothesis that no untoward racial bias is present in this ML model. Were this ML decision-support model put into production use to guide prescribing, it is unlikely that it would manifest racially discriminatory or unjust recommendations.

For the medulloblastoma follow-up example, **Table 6** likewise shows features of the 10 top-ranked models generated by BMA MCMC sampling, together with the cumulative inclusion probability for each feature summed over the models evaluated.

Variable	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	Incl. Prob.
Severe CHF	X	X	X	X	X	X	X	X	X	X	1.00
Asthma	X	X	X	X	X	X	X	X	X	X	1.00
Bradycardia	X	X	X	X	X	X	X	X		X	1.00
Heart block	X	X	X	X	X	X	X				1.00
Pacemaker	X	X	X	X	X					X	0.91
Hypotension (SBP < 105 mm)	X	X	X	X	X						1.00
Polypharmacy (Nmeds > 10)	X	X		X		X		X	X		0.88
Pressors or inotrope	X	X	X		X	X	X			X	0.79
IABP or VAD	X		X	X		X		X	X		0.62
Urgent-emergent	X	X			X		X	X		X	0.54
Age	X				X	X			X		0.21
Race				X				X			0.13
Posterior Model Probability	0.021	0.012	0.011	0.010	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	

Table 5. BMA of discharge beta-blocker utilization, post-CAB.

Variable	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	Incl. Prob.
Clinical trial	X	X	X	X	X	X	X	X	X	X	1.00
Age	X	X	X	X	X	X	X	X	X		0.99
Prior recurrence	X	X	X	X	X		X	X		X	0.97
Renal impairment	X	X	X	X		X		X	X		0.96
High-risk histology	X	X	X	X	X	X		X			0.94
SHH/WNT genomics	X	X	X		X	X	X		X	X	0.70
Tumor extent	X	X	X		X	X		X	X	X	0.85
PFS	X	X		X	X		X	X		X	0.46
^{99m}Tc scan	X	X	X			X	X				0.32
mIGB scan	X	X	X		X		X	X		X	0.29
Race	X	X	X	X	X	X		X		X	0.58
Public payor	X		X				X		X		0.17
Posterior Model Probability	0.032	0.024	0.017	0.011	0.010	<0.01	<0.01	<0.01	<0.01	<0.01	

Table 6. BMA of serial MRI utilization, post-medulloblastoma resection.

With regard to repeat MRI in follow-up of pediatric medulloblastoma, Bayesian model averaging yields evidence that some models that include race have higher Posterior Model probability (PMP) than models that omit race as a feature, and race exhibits relatively high inclusion probability among the 1000 top-ranked models. These findings are consistent with the results of CMH and beta regression. The evidence suggests that the ML model manifests biases which, if put into production use to guide prescribing, may reproduce or exaggerate disparities that were present in the historical observational data from which the ML model was learned.

5. Discussion

Machine-learning methods are finding increasing application to guide decision-making, including decisions that arise in health care. ML decision-support guidance can have important consequences on outcomes, including employment [50, 51, 70, 71], banking [10], predictive policing and law enforcement [2, 11, 12, 21, 49, 72–74], and criminal sentencing [48]. Recently, growing attention has focused on the potential that machine-learning might learn unjust, unfair, or discriminatory representations from observational data and inadvertently

perpetuate or propagate injustices that are manifested in the historical data that are utilized to train the machine-learning models [2, 7]. Despite the increasing attention to this issue, as yet it is unclear whether the goals of fairness and accuracy in ML are conflicting goals [5, 19, 24]. In that connection, the impact of race/ethnicity on health services access, long-term risk factor control, and cardiovascular outcomes among patients has been the subject of intensive study for decades [75–83]. However, significant disparities in cardiovascular management have received less attention [84–89]. Similarly, the current literature has directed scant attention to disparities in cancer care subsequent to diagnosis. In the present era of artificial intelligence, Big Data, and machine-learning, it is a priority that ML-based decision-support tools not manifest untoward disparities. Sensitive methods having statistical power adequate to detect disparity are essential to achieving this goal. Moreover, it is important that such methods be aligned with generally-accepted governance practices in the courts and regulatory agencies.

The present work sets forth a three-pronged approach for ascertaining the presence or absence of disparity in ML models, by race, age, gender, or other attributes. We sought to discover strengths and limitations of methods for detecting unfairness in ML-model-guided decision-support and, when unfairness is identified, discovering the sources and magnitudes of the disparate effects. In health services, numerous clinical contexts and models and treatment use-cases merit such analyses. However, for simplicity we selected two contexts in which strong consensus does exist regarding what the preferred treatment should be, and in which the consensus has prevailed and remained constant for a sufficient period of time, such that observational data are available for analysis and such that minimal change in the consensus has occurred during the time period for which data are available for analysis. We selected cardiac care and cancer care contexts in which disparities with respect to race are feasible evaluate.

Other factors such as socioeconomic status and health services access patterns remain to be studied. The frequency and tenure of accessing the health system are confounded by race and socioeconomic factors. Patients' frequency and tenure also influence what medications patients have been prescribed previously [76], including some medications that may have been discontinued or substituted due to side-effects or non-efficacy reasons, events that influence subsequent considerations for devising or adjusting the patients' medications regimen when new circumstances arise. Nonetheless, we examined one example intervention that has been regarded as 'standard if care' for a long time and whose marginal cost in the U.S. context is so small as to be negligible (beta-blocker medications at hospital discharge post-CAB) and another intervention whose marginal cost in the U.S. is significant (serial MRI exams of head and spinal cord).

Beta-blockers have been found to be important in the treatment of myocardial infarction and in coronary artery bypass surgery in that they have been shown to decrease mortality. Their use post-CAB has been standard care for many years [47], conditioned on the absence of significant clinical contraindications to their use in a particular patient. However, prescribing a beta-blocker at the time of discharge from hospital post-CAB remains less consistent than it should be. Of note is that most beta-blocker medications are extensively metabolized by the liver (esp. CYP2D6) and are affected by liver function. Indeed, the concomitant use of

CYP2D6-inhibiting medications or the presence of liver disease may contraindicate or restrict the use of beta-blockers. Our ML modeling process determined the statistical significance and retention of AST and the AST/ALT ratio in the ML predictive model, consistent with this anticipated relevance of liver function to prescribers' decision-making, recapitulated in the model. However, alcohol use, hepatitis, non-alcoholic steatosis, cirrhosis, and other liver conditions are known to exhibit racial imbalance. Slightly elevated prevalence of cirrhosis in has been reported in the U.S. black population (viz., QT_c prolongation and the risk of ventricular arrhythmias, see [40]). At the outset of the present study, we were concerned that such imbalances might confound the ML modeling process or give rise to an ML model that could exacerbate under-prescribing of beta-blockers to black individuals.

By using the Cochran-Mantel-Haenszel test, beta regression, and Bayesian Model Averaging, not only was no untoward racial disparity found in the post-CAB cohort, but, with regard to the likelihood of receiving standard-of-care discharge beta-blocker after CAB, there was unexpectedly a slight benefit associated with black race.

By using the Cochran-Mantel-Haenszel test, a statistically significant and unexpected racial disparity was detected in the medulloblastoma ML model in regard to serial repeat MRI exams following initial cancer treatment. This was corroborated by beta regression and confirmed by Bayesian model averaging analysis, wherein many BMA-generated models retained race as a statistically significant predictor of serial MRI utilization. Potential reasons for the disparity are the subject of ongoing study.

Presently, we explicitly exclude race as an input variable from both of the ML models discussed as examples above, as a matter of assuring that the models will not perpetuate clinical differences in utilization rates associated with race or ethnicity manifested in the observational data used to train and validate the models. Naturally, race is only one factor that might be considered as a basis for potential unfairness. Attention should be directed also to other attributes that are candidate predictors in ML models, such as age, gender, chronicity/tenure or survival phase, payor class, or previous exposure to treatments or procedures that themselves might be subject to disparities, inequitable rationing, or unjust differential access or provisioning rates between groups. Confounding may arise from other factors [90–109], such as the vigor or effectiveness with which informed consent is sought by the treating physicians. Such confounding may be affected by racial or cultural differences between the family and the person performing the consenting process. This merits ongoing evaluation by model developers and model users, to insure that good and fair ML models are not erroneously disparaged or rejected for invalid reasons.

As revealed in the example of medulloblastoma treatment follow-up, quantitative Bayesian and frequentist surveillance for potential model unfairness may detect phenomena that are not evidence of injustice per se but instead reflect cultural, educational, religious/spiritual, coping style, family structure, economic, comorbid anxiety/depression rates, rurality, inability to leave work, or other underlying sociodemographic differences. Such differences in decision-making are worthy foci of bioethical, epidemiological, and other evaluations, but are not necessarily differences that merit sanction or suppression. Autonomy of patients' and families' decision-making must be respected. Thus, fair, equitable, nondiscriminatory offering of

options and access to services to all does not compel equal utilization of services by all [110]. Nonetheless, financial barriers to care may prevent minority and underserved populations from accessing follow-up care at rates commensurate with other groups. Enhancing insurance coverage or addressing out-of-pocket costs may help address financial barriers to follow-up care, including repeat screening to detect recurrence or progression.

Compared to CMH and beta regression, BMA is able to achieve adequate power with smaller sample sizes. Moreover, BMA does not have the odds ratio homogeneity, parametric distributional, or other disadvantages of CMH or beta regression. Our BMA analytical approach meets the primary goals for defining a statistical approach to assessing fairness of ML models. Specifically:

1. It captures information from the endpoint scale on the interval [0,1);
2. It provides an interpretation that is readily understood;
3. It has power at least equivalent to the CMH test or beta regression;
4. It avoids assumptions in the calculation of the significance of treatment differences; and
5. The interpretation is based on the same foundation as the calculation of the p-value.

We suggest that that this approach is superior to the stratified dichotomous approach as it captures the entire spectrum of the outcome scale, and therefore will be generally more powerful. While it remains valuable to use a combination of two or more methods (including frequentist methods, such as CMH and beta regression) in a correlative manner to insure consistent determination of fairness of ML models, BMA has become for us a preferred component of fairness testing owing to its modestly greater statistical power when some strata have small size or there is marked unbalancing among strata. Bayesian methods, including BMA, are essential components of auditing processes and policy-setting processes for ML decision-support models, and are valuable adjuncts to conventional frequentist methods, which are less well-suited to the combinatorial challenges of high dimensionality in model variables selection in the Big Data era.

In summary, we propose that these frequentist and Bayesian methods, including BMA, may be valuable for other outcomes types and other contexts and use-cases, to detect disparities in a fashion similar to how statistical tests have historically been used in the courts and in public policy-making and regulation [52, 111]. Based on our success with the present example use-case, we particularly recommend that BMA may be valuable for other use-cases in health services-related ML modeling, to determine the covariate sources of ML-model-based decision-support disparities that are discovered, to measure the magnitudes of such effects, and to perform model-curation quality assurance so as to insure that such disparities can be eliminated [18] or kept to minimum levels. Such methods may help to promote and quantify algorithmic fairness [8–31], assist in proper governance of ML-based decision-support tools, and insure that ML modeling does not inadvertently learn and replicate unfair practices that are extant in observational datasets that are mined, thereby avoiding perpetuation of injustice by artificial intelligence or cognitive computing. These methods appear to be adequately sensitive

and effective in terms of statistical power for cohort sizes such as are practically available. The frequentist methods have the advantage of general acceptance in the public sector and a long history of use in the courts and in regulatory settings. However, they are not well-suited to Big Data with high dimensionality and significant missingness rates for individual predictor variables. By contrast, BMA does not yet have a history of use in the courts or in other public-policy or regulatory settings. Nonetheless, confirmation by BMA of the statistically negligible role of race in our post-CAB cohort and a likely significant role of race in our medulloblastoma cohort suggests that BMA should be an important addition to the toolbox supporting fairness-assurance of ML models in these and similar contexts and can also help courts and regulators ascertain fairness of decision-support models in actual application. Correspondingly, BMA can help model developers to defend against allegations of unfairness as they arise.

Author details

Douglas S. McNair

Address all correspondence to: dmcnair@cerner.com

Cerner Corporation, Kansas City, USA

References

[1] Collmann J, Matei S, editors. Ethical Reasoning in Big Data: An Exploratory Analysis. Berlin: Springer Verlag; 2016. p. 192

[2] Coglianese C, Lehr D. Regulating by robot: Administrative decision-making in the machine-learning era. Georgetown Law Journal. 28-Feb-2017. Available: http://scholarship.law.upenn.edu/faculty_scholarship/1734/

[3] Fox M. Technology is a marvel – Now let's make it moral. The Guardian, April 10, 2017. Available: https://www.theguardian.com/commentisfree/2017/apr/10/ethical-technology-women-britain-internet

[4] Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. Proceedings of the 3rd ACM Innovations in Theoretical Computer Science Conference. 2012;**2012**:214-226

[5] Joseph M, Kearns M, Morgenstern J, Roth A. Fairness in learning: Classic and contextual bandits. arXiv preprint, 2016. Available: https://arxiv.org/pdf/1605.07139

[6] Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C. Learning fair representations. Proceedings of the 30th International Conference on Machine Learning. 2013;**2013**:325-333

[7] Barocas S, Selbst A. Big Data's disparate impact. California Law Review 2016;**104**:671-733. Available: http://ssrn.com/abstract=2477899

[8] Bechavod Y, Ligett K. Learning fair classifiers: A regularization-inspired approach. KDD '17, Halifax 2017, ACM. Available: https://arxiv.org/pdf/1707.00044.pdf

[9] Bunnik A, Cawley A, Mulqueen M, Zwitter A, editors. Big Data Challenges: Society, Security, Innovation and Ethics. London: Palgrave Macmillan; 2016. p. 140

[10] Byrnes N. Artificial intolerance. MIT Technology Review, March 28 2016. Available: https://www.technologyreview.com/s/600996/artificial-intolerance/

[11] Chouldechova A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. arXiv preprint, 2016. Available: https://arxiv.org/pdf/1610.07524

[12] Chouldechova A, G'Sell M. Fairer and more accurate, but for whom? KDD '17, Halifax 2017, ACM. Available: https://arxiv.org/pdf/1707.00046.pdf

[13] Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A. Algorithmic decision-making and the cost of fairness. Stanford Working Paper, 18-FEB-2017, Available: https://arxiv.org/pdf/1701.08230

[14] Béranger J. Big Data and Ethics: The Medical Datasphere. New York: ISTE Press/Elsevier; 2016. p. 300

[15] Davis K. Ethics of Big Data: Balancing Risk and Innovation. Sebastopol, CA: O'Reilly Media; 2012. p. 82

[16] Berk R, Heidari H, Jabbari S, Joseph M, Kearns M, Morgenstern J, Neel S, Roth A. A convex framework for fair regression. KDD '17, ACM. Available: https://arxiv.org/pdf/1706.02409.pdf

[17] FAT/ML (Fairness, Accountability, and Transparency in Machine Learning) Available: http://www.fatml.org

[18] Feldman M, Friedler S, Moeller J, Scheidegger C, Venkatasubramanian S. Certifying and removing disparate impact. arXiv preprint, 2015. Available: https://arxiv.org/abs/1412.3756

[19] Fish B, Kun J, Lelkes A. A Confidence-Based Approach for Balancing Fairness and Accuracy. SIAM International Conference on Data Mining, 2016. Available: http://homepages.math.uic.edu/~bfish3/sdm_2016.pdf

[20] Francez N. Fairness. Berlin: Springer Verlag; 1986. p. 298

[21] Guinn C. Big data algorithms can discriminate, and it's not clear what to do about it. The Conversation blog, Aug 13, 2015. Available: http://theconversation.com/big-data-algorithms-can-discriminate-and-its-not-clear-what-to-do-about-it-45849

[22] Hajian S, Bonchi F, Castillo C. Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining. KDD '16, August 2016. San Francisco: ACM. ISBN 978-1-4503-4232-2/16/08. DOI: 10.1145/2939672.2945386

[23] Hodson H. No one in control: The algorithms that run our lives. New Scientist, 05-FEB-2015

[24] Jabbari S, Joseph M, Kearns M, Morgenstern J, Roth A. Fair learning in Markovian environments. arXiv preprint, 2016. Available: https://arxiv.org/pdf/1611.03071

[25] Mittelstadt B, Floridi L, editors. The Ethics of Biomedical Big Data. Berlin: Springer Verlag; 2016. p. 480

[26] O'Neil C. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown; 2016. p. 300

[27] Robertson J, Webb W. Cake-Cutting Algorithms: Be Fair if you can. Boca Raton: CRC Press; 1998. p. 300

[28] Simoiu C, Corbett-Davies S, Goel S. The problem of infra-marginality in outcome tests for discrimination. arXiv preprint, 2017. Available: https://arxiv.org/pdf/1607.05376

[29] Skirpan M, Gorelick M. The authority of 'fair' in machine learning. KDD '17, Halifax 2017, ACM. Available: https://arxiv.org/pdf/1706.09976.pdf

[30] Veal M. Logics and practices of transparency and opacity in real-world applications of public sector machine learning. KDD '17, Halifax 2017, ACM. Available: https://arxiv.org/pdf/1706.09249.pdf

[31] Zhang Z, Neill D. Identifying significant predictive bias in classifiers. KDD '17, Halifax 2017, ACM. Available: https://arxiv.org/pdf/1611.08292.pdf

[32] Steinberg C, Padfield GJ, Al-Sabeq B, Adler A, Yeung-Lai-Wah JA, Kerr CR, Deyell MW, Andrade JG, Bennett MT, Yee R, Klein GJ, Green M, Laksman ZW, Krahn AD, Chakrabarti S. Experience with bisoprolol in long-QT1 and long-QT2 syndrome. Journal of Interventional Cardiac Electrophysiology. 2016;**47**(2):163-170

[33] Servaes S, Epelman M, Pollock A, Shekdar K. Pediatric malignancies: Synopsis of current imaging techniques. In: Blake M, Kalra M, editors. Imaging in Oncology. New York: Springer; 2008. pp. 469-492

[34] Tibshirani R. Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society, Series B. 1996;**58**(1):267-288

[35] Delker E, Brown Q, Hasin DS. Alcohol consumption in demographic subpopulations: An epidemiologic overview. Alcohol Research: Current Reviews. 2016;**38**:7-15

[36] Hughson MD, Puelles VG, Hoy WE, Douglas-Denton RN, Mott SA, Bertram JF. Hypertension, glomerular hypertrophy and nephrosclerosis: The effect of race. Nephrology, Dialysis, Transplantation. 2014;**29**:1399-1409

[37] Klous S, Wielaard N. We Are Big Data: The Future of the Information Society. New York: Atlantis Press; 2016. p. 300

[38] Na L et al. Disparities in receipt of recommended care among younger versus older Medicare beneficiaries: A cohort study. BMC Health Services Research. 2017;**17**:241-253

[39] Sajja K, Mohan DP, Rockey DC. Age and ethnicity in cirrhosis. Journal of Investigational Medicine. 2014;**62**:920-926

[40] Tuttolomondo A, Buttà C, Casuccio A, Di Raimondo D, Serio A, D'Aguanno G, Pecoraro R, Renda C, Giarrusso L, Miceli G, Cirrincione A, Pinto A. QT indexes in cirrhotic patients: Relationship with clinical variables and potential diagnostic predictive value. Archives of Medical Research. 2015;**46**:207-213

[41] Valles S. Heterogeneity of risk within racial groups: A challenge for public health programs. Preventive Medicine. 2012;**55**:405-408

[42] Yu Q, Fan Y, Wu X. General multiple mediation analysis with an application to explore racial disparity in breast cancer survival. Journal of Biometrics & Biostatistics. 2014;**5**: 189-196

[43] Yu Q, Scribner RA, Leonardi C, Zhang L, Park C, Chen L, Simonsen NR. Exploring racial disparity in obesity: A mediation analysis considering geo-coded environmental factors. Spatial & Spatio-temporal Epidemiology. 2017;**21**:13-23

[44] Schonberger R, Gilbertsen T, Dai F. The problem of controlling for imperfectly measured confounders on dissimilar populations: A database simulation study. Journal of Cardiothoracic and Vascular Anesthesia. 2014;**28**:247-254

[45] Amsterdam EA, Wenger NK, Brindis RG, Casey DE Jr, Ganiats TG, Holmes DR Jr, Jaffe AS, Jneid H, Kelly RF, Kontos MC, Levine GN, Liebson PR, Mukherjee D, Peterson ED, Sabatine MS, Smalling RW, Zieman SJ. 2014 AHA/ACC guideline for the management of patients with non–ST-elevation acute coronary syndromes. Journal of the American College of Cardiology 2014;**64**:e139-e228

[46] Boudonas G. β-blockers in coronary artery disease management. Hippokratia. 2010;**14**: 231-235

[47] Khan M. Cardiac Drug Therapy. 7th ed. Totawa, NJ: Humana Press; 2007. p. 420

[48] Barry-Jester A. The new science of sentencing. The Marshall Project, 2015. Available: https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing

[49] Benforado A. Unfair: The New Science of Criminal Injustice. New York: Crown; 2015. p. 300

[50] Gastwirth J. Statistical methods for analyzing claims of employment discrimination. Industrial & Labor Relations Review. 1984;**38**:75-86

[51] Gastwirth J et al. Statistical methods for assessing the fairness of the allocation of shares in initial public offerings. Law Probability & Risk. 2005. DOI: 10.1093/lpr/mgi012

[52] Kadane J. Statistics in the Law: A Practitioner's Guide, Cases, and Materials. Oxford: Oxford University Press; 2008. p. 472

[53] Paolino. *maximum* Likelihood estimation of models with beta-distributed dependent variables. Political Analysis. 2001;**9**:325-346

[54] Smithson M, Verkuilen J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychological Methods. 2006;**11**:54-71

[55] Ferrari S, Cribari-Neto F. Beta regression for modelling rates and proportions. Journal of Applied Statistics. 2004;**31**:799-815

[56] Cribari-Neto F, Zeileis A. Beta regression in R. Journal of Statistical Software. 2010;**34**:1-24

[57] Smithson M, Deady S, Gracik L. Guilty, not guilty, or…?: Multiple options in jury verdict choices. Journal of Behavioral Decision Making. 2007;**20**:481-498

[58] Hubben G, Bishai D, Pechlivanoglou P, Cattelan AM, Grisetti R, Facchin C, Compostella FA, Bos JM, Postma MJ, Tramarin A. The societal burden of HIV/AIDS in northern Italy: An analysis of costs and quality of life. AIDS Care: Psychological and Socio-Medical Aspects of AIDS/HIV. 2008;**20**:449-455

[59] Basu A, Manca A. Regression estimators for generic health-related quality of life and quality-adjusted life years. Medical Decision Making. 2012;**32**:56-69

[60] Ando T. Bayesian Model Selection and Statistical Modeling. Boca Raton: CRC Press; 2010. p. 300

[61] Bayarri MJ, Berger JO, Forte A, Garcia-Donato G. Criteria for Bayesian model choice with application to variable selection. Ann. Statist. 2012;**40**:1550-1577

[62] Claeskens G, Hjort N. Model Selection and Model Averaging. Cambridge: Cambridge University Press; 2008. p. 332

[63] Eicher T, Papageorgiou C, Raftery A. Determining growth determinants: Default priors and predictive performance in Bayesian model averaging. Journal of Applied Econometrics. 2011;**26**(1):30-55

[64] Garcia-Donato G, Forte A. R package BayesVarSel, Available: https://cran.r-project.org/package=BayesVarSel

[65] Feldkircher M, Zeugner S. Benchmark Priors Revisited: On Adaptive Shrinkage and the Supermodel Effect in Bayesian Model Averaging. IMF Working Paper, WP/09/202, 2009. DOI:10.5089/9781451873498.001

[66] Fernandez C, Ley E, Steel MF. Benchmark priors for Bayesian model averaging. Journal of Econometrics. 2001;**100**:381-427

[67] Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: A tutorial. Statistical Science. 1999;**14**(4):382-417

[68] Zellner A. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Zellner A, editor. Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. Edward Elgar Publishing: London; 1986. pp. 389-399

[69] Liang F, Paulo R, Molina G, Clyde MA, Berger JO. Mixtures of g priors for Bayesian variable selection. Journal of the American Statistical Association. 2008;**103**:410-423

[70] Bertrand M, Mullainathan S. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. The American Economic Review. 2004;**94**:991-1013

[71] Miller C. Can an algorithm hire better than a human? The New York Times, 25-Jun-2015. Available: http://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html

[72] Goel S, Rao J, Shroff R. Personalized risk assessments in the criminal justice system. The American Economic Review. 2016;**106**:119-123

[73] Goel S, Rao J, Shroff. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. Annals of Applied Statistics. 2016;**10**:365-394

[74] Lum K, Isaac W. To predict and serve? Significance. 2016;**13**:14-19

[75] Amini A, Yeh N, Jones BL, Bedrick E, Vinogradskiy Y, Rusthoven CG, Amini A, Purcell WT, Karam SD, Kavanagh BD, Guntupalli SR, Fisher CM. Perioperative mortality in non-elderly adult patients with cancer: A population-based study evaluating health care disparities in the united states according to insurance status. American Journal of Clinical Oncology. 2016 Jun 8. DOI: 10.1097/COC.0000000000000306

[76] Beohar N et al. Race/ethnic disparities in risk factor control and survival in the bypass angioplasty revascularization investigation 2 diabetes (BARI-2D) trial. American Journal of Cardiology. 2013;**112**:1298-1305

[77] Buja A, Boemo DG, Furlan P, Bertoncello C, Casale P, Baldovin T, Marcolongo A, Baldo V. Tackling inequalities: Are secondary prevention therapies for reducing post-infarction mortality used without disparities? European Journal of Preventive Cardiology. 2014;**21**:222-230

[78] Butwick A, Blumenfeld YJ, Brookfield KF, Nelson LM, Weiniger CF. Racial and ethnic disparities in mode of anesthesia for cesarean delivery. Anesthesia & Analgesia. 2016;**122**:472-479

[79] Cheng E, Declercq ER, Belanoff C, Iverson RE, McCloskey L. Racial and ethnic differences in the likelihood of vaginal birth after caesarean delivery. Birth. 2015;**42**:249-253

[80] Efird J, Griffin WF, Sarpong DF, Davies SW, Vann I, Koutlas NT, Anderson EJ, Crane PB, Landrine H, Kindell L, Iqbal ZJ, Ferguson TB, Chitwood WR, Kypson AP. Increased long-term mortality among black CABG patients receiving preoperative inotropic agents. International Journal of Environmental Research and Public Health. 2015;**12**: 7478-7490

[81] Efird J, Gudimella P, O'Neal WT, Griffin WF, Landrine H, Kindell LC, Davies SW, Sarpong DF, O'Neal JB, Crane P, Nelson M, Ferguson TB, Chitwood WR, Kypson AP, Anderson EJ. Comparison of risk of atrial fibrillation in black versus white patients after coronary artery bypass grafting. The American Journal of Cardiology. 2016;**117**:1095-1100

[82] Efird J, Kiser AC, Crane PB, Landrine H, Kindell LC, Nelson MA, Jindal C, Sarpong DF, Griffin WF, Ferguson TB, Chitwood WR, Davies SW, Kypson AP, Gudimella P, Anderson EJ. Perioperative inotrope therapy and atrial fibrillation following coronary artery bypass graft surgery: Evidence of a racial disparity. Pharmacotherapy. 2017;**37**: 297-304

[83] Shiels M, Chernyavskiy P, Anderson WF, Best AF, Haozous EA, Hartge P, Rosenberg PS, Thomas D, Freedman ND, Berrington de Gonzalez A. Trends in premature mortality in the USA by sex, race, and ethnicity from 1999 to 2014: An analysis of death certificate data. Lancet. 2017;**389**:1043-1054

[84] Brown C, Ross L, Lopez I, Thornton A, Kiros GE. Disparities in the receipt of cardiac revascularization procedures between blacks and whites: An analysis of secular trends. Ethnic Disparities. 2008;**18**(2 Suppl 2):112-117

[85] Dimick J, Ruhter J, Sarrazin MV, Birkmeyer JD. Black patients more likely than whites to undergo surgery at low-quality hospitals in segregated regions. Health Affairs (Millwood). 2013;**32**:1046-1053

[86] Mehta RH, Shahian DM, Sheng S, O'Brien SM, Edwards FH, Jacobs JP, Peterson ED. Association of hospital and physician characteristics and care processes with racial disparities in procedural outcomes among contemporary patients undergoing coronary artery bypass grafting surgery. Circulation. 2016;**133**:124-130

[87] Nallamothu B, Lu X, Vaughan-Sarrazin MS, Cram P. Coronary revascularization at specialty cardiac hospitals and peer general hospitals in black Medicare beneficiaries. Circulation. Cardiovascular Quality and Outcomes. 2008;**1**:116-122

[88] O'Neal W, Efird JT, Davies SW, O'Neal JB, Griffin WF, Ferguson TB, Chitwood WR, Kypson AP. Discharge β-blocker use and race after coronary artery bypass grafting. Frontiers in Public Health. 2014;**2**:94-99

[89] Rangrass G, Ghaferi AA, Dimick JB. Explaining racial disparities in outcomes after cardiac surgery: The role of hospital quality. JAMA Surgery. 2014;**149**:223-227

[90] Best AL, Alcaraz KI, McQueen A, Cooper DL, Warren RC, Stein K. Examining the mediating role of cancer-related problems on spirituality and self-rated health among African American cancer survivors: A report from the American Cancer Society's studies of cancer survivors-II. Psycho-Oncology. 2015;**24**:1051-1059. DOI: 10.1002/pon.3720

[91] Bromley EG, May FP, Federer L, Spiegel BM, van Oijen MG. Explaining persistent under-use of colonoscopic cancer screening in African Americans: A systematic review. Preventive Medicine 2015;**71**:40-48. doi: 10.1016/j.ypmed.2014.11.022

[92] Christman LK, Abernethy AD, Gorsuch RL, Brown A. Intrinsic religiousness as a mediator between fatalism and cancer-specific fear: Clarifying the role of fear in prostate cancer screening. Journal of Religion and Health. 2014;**53**(3):760-772. DOI: 10.1007/s10943-012-9670-1

[93] Davis JL, Bynum SA, Katz RV, Buchanan K, Green BL. Sociodemographic differences in fears and mistrust contributing to unwillingness to participate in cancer screenings. Journal of Health Care for the Poor and Underserved. 2012;**23**(4 Suppl):67-76. DOI: 10.1353/hpu.2012.0148

[94] Glickman SW, Anstrom KJ, Lin L, Chandra A, Laskowitz DT, Woods CW, Freeman DH, Kraft M, Beskow LM, Weinfurt KP, Schulman KA, Cairns CB. Challenges in enrollment of minority, pediatric, and geriatric patients in emergency and acute care clinical research. Annals of Emergency Medicine. 2008;**51**:775-780. DOI: 10.1016/j.annemergmed.2007.11.002

[95] Gourlay ML, Lewis CL, Preisser JS, Mitchell CM, Sloane PD. Perceptions of informed decision making about cancer screening in a diverse primary care population. Family Medicine. 2010;**42**(6):421-427

[96] Hamilton JB, Best NC, Galbraith KV, Worthy VC, Moore LT. Strategies African-American cancer survivors use to overcome fears and fatalistic attitudes. Journal of Cancer Education. 2015;**30**(4):629-635. DOI: 10.1007/s13187-014-0738-3

[97] Koch C, Li L, Kaplan GA, Wachterman J, Shishehbor MH, Sabik J, Blackstone EH. Socioeconomic position, not race, is linked to death after cardiac surgery. Circulation. Cardiovascular Quality and Outcomes. 2010;**3**:267-276

[98] Miller SJ, Iztkowitz SH, Redd WH, Thompson HS, Valdimarsdottir HB, Jandorf L. Colonoscopy-specific fears in African Americans and Hispanics. Behavioral Medicine. 2015;**41**(2):41-48. DOI: 10.1080/08964289.2014.897930

[99] Nagelhout E, Comarell K, Samadder NJ, Wu YP. Barriers to colorectal cancer screening in a racially diverse population served by a safety-net clinic. Journal of Community Health. 2017;**42**(4):791-796. DOI: 10.1007/s10900-017-0319-6

[100] Owens OL, Jackson DD, Thomas TL, Friedman DB, Hébert JR. African American men's and women's perceptions of clinical trials research: Focusing on prostate cancer among a high-risk population in the south. Journal of Health Care for the Poor and Underserved. 2013;**24**(4):1784-1800. DOI: 10.1353/hpu.2013.0187

[101] Palmer NR, Weaver KE, Hauser SP, Lawrence JA, Talton J, Case LD, Geiger AM. Disparities in barriers to follow-up care between African American and white breast cancer survivors. Supportive Care in Cancer. 2015;**23**(11):3201-3209. DOI: 10.1007/s00520-015-2706-9

[102] Pandya D, Patel S, Ketchum NS, Pollock BH, Padmanabhan S. A comparison of races and leukemia subtypes among patients in different cancer survivorship phases. Clinical Lymphoma, Myeloma & Leukemia. 2011;**11**(Suppl 1):S114-S118. DOI: 10.1016/j.clml.2011.05.036

[103] Pittman LJ. A thirteenth amendment challenge to both racial disparities in medical treatment and improper physicians' informed consent disclosures. Saint Louis University School of Law. 2003;**48**(1):131-189

[104] Shaw MG, Morrell DS, Corbie-Smith GM, Goldsmith LA. Perceptions of pediatric clinical research among African American and Caucasian parents. Journal of the National Medical Association. 2009;**101**(9):900-907

[105] Shepperd JA, Howell JL, Logan H. A survey of barriers to screening for oral cancer among rural black Americans. Psycho-Oncology. 2014;**23**(3):276-282. DOI: 10.1002/pon.3415

[106] Taylor M, Sun AY, Davis G, Fiuzat M, Liggett SB, Bristow MR. Race, common genetic variation, and therapeutic response disparities in heart failure. JACC Heart Failure. 2014;**2**:561-572

[107] Taylor TR, Huntley ED, Sween J, Makambi K, Mellman TA, Williams CD, Carter-Nolan P, Frederick W. An exploratory analysis of fear of recurrence among African-American breast cancer survivors. International Journal of Behavioral Medicine. 2012;**19**(3):280-287. DOI: 10.1007/s12529-011-9183-4

[108] Torke AM, Corbie-Smith GM, Branch WT. African American patients' perspectives on medical decision making. Archives of Internal Medicine. 2004;**164**(5):525-530

[109] Vrinten C, Wardle J, Marlow LA. Cancer fear and fatalism among ethnic minority women in the United Kingdom. British Journal of Cancer. 2016;**114**(5):597-604. DOI: 10.1038/bjc.2016.15

[110] Macklin R. Ethical relativism in a multicultural society. Kennedy Institute of Ethics Journal. 1998;**8**(1):1-22

[111] Vuolo M, Uggen C, Lageson S. Statistical power in experimental audit studies: Cautions and calculations for matched tests with nominal outcomes. Sociological Methods & Research. 2016;**45**:260-303