Information Systems Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

When Algorithmic Predictions Use Human-Generated Data: A Bias-Aware Classification Algorithm for Breast Cancer Diagnosis

Mehmet Eren Ahsen, Mehmet Ulvi Saygi Ayvaci, Srinivasan Raghunathan

To cite this article:
Mehmet Eren Ahsen, Mehmet Ulvi Saygi Ayvaci, Srinivasan Raghunathan (2019) When Algorithmic Predictions Use Human-Generated Data: A Bias-Aware Classification Algorithm for Breast Cancer Diagnosis. Information Systems Research 30(1):97-116. https://doi.org/10.1287/isre.2018.0789

http://pubsonline.informs.org/journal/isre/

INFORMATION SYSTEMS RESEARCH
Vol. 30, No. 1, March 2019, pp. 97–116
ISSN 1047-7047 (print), ISSN 1526-5536 (online)

When Algorithmic Predictions Use Human-Generated Data: A Bias-Aware Classification Algorithm for Breast Cancer Diagnosis

Mehmet Eren Ahsen,[a] Mehmet Ulvi Saygi Ayvaci,[b] Srinivasan Raghunathan[b]

[a] Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York 10029; [b] Jindal School of Management, University of Texas at Dallas, Richardson, Texas 75080

Contact: mehmet.erenahsen@mssm.edu (MEA); mehmet.ayvaci@utdallas.edu, http://orcid.org/0000-0001-6997-1639 (MUSA); sraghu@utdallas.edu (SR)

Received: August 9, 2016
Revised: June 23, 2017; December 13, 2017
Accepted: February 26, 2018
Published Online in Articles in Advance: December 20, 2018

https://doi.org/10.1287/isre.2018.0789

Abstract. When algorithms use data generated by human beings, they inherit the errors stemming from human biases, which likely diminishes their performance. We examine the design and value of a bias-aware linear classification algorithm that accounts for bias in input data, using breast cancer diagnosis as our specific setting. In this context, a referring physician makes a follow-up recommendation to a patient based on two inputs: the patient's clinical-risk information and the radiologist's mammogram assessment. Critically, the radiologist's assessment could be biased by the clinical-risk information, which in turn can negatively affect the referring physician's performance. Thus, a bias-aware algorithm has the potential to be of significant value if integrated into a clinical decision support system used by the referring physician. We develop and show that a bias-aware algorithm can eliminate the adverse impact of bias if the error in the mammogram assessment due to radiologist's bias has no variance. On the other hand, in the presence of error variance, the adverse impact of bias can be mitigated, but not eliminated, by the bias-aware algorithm. The bias-aware algorithm assigns less (more) weight to the clinical-risk information (radiologist's mammogram assessment) when the mean error increases (decreases), but the reverse happens when the error variance increases. Using point estimates obtained from mammography practice and the medical literature, we show that the bias-aware algorithm can significantly improve the expected patient life years or the accuracy of decisions based on mammography.

History: Giri Kumar Tayi, Senior Editor; Sam Ransbotham, Associate Editor.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/isre.2018.0789.

Keywords: algorithms • decision support systems • classification • bias • mammography • medical decision making

> Algorithms that learn from human decisions will also learn human mistakes, such as overtesting and overdiagnosis, failing to notice people who lack access to care, undertesting those who cannot pay, and mirroring race or gender biases. Ignoring these facts will result in automating and even magnifying problems in our current health system. Noticing and undoing these problems requires a deep familiarity with clinical decisions and the data they produce–a reality that highlights the importance of viewing algorithms as thinking partners, rather than replacements, for doctors.
>
> —Obermeyer and Lee (2017, p. 1211)

1. Introduction

Major advancements in information technology (IT) in the last decade have produced an unprecedented rise in readily available and complex data sets. Such ubiquity of data has facilitated an increased use of algorithms—analytical tools that use available data and produce an output (such as a prediction)—by the decision makers in social and economic transactions. The pervasiveness of data promises significant opportunities to improve decision making in every conceivable domain with the help of the algorithmic approach (Agarwal and Dhar 2014), and efforts are already underway in domains ranging from travel-pricing forecasts (Elliott 2017) to the prediction of recidivism (Brennan et al. 2009) to diagnostic decisions in medicine (Chen and Asch 2017).

An important advantage of the algorithmic approach to decision making is that it may not suffer from the cognitive limitations or biases that are typically associated with human decision making. However, when the input data used by the algorithms are generated by human beings, even algorithms become susceptible to human biases. The bias in input data would likely diminish the algorithm's performance. For a specific example in healthcare, consider an algorithm that predicts the probability of breast cancer for a patient

based on the patient's clinical characteristics and the radiologist's assessment of the patient's mammogram. However, the radiologist often accesses the patient's clinical characteristics while providing her mammogram assessment, and her assessment is influenced by the patient's clinical characteristics (Carney et al. 2012). The phenomenon—the use of human-generated decisions as inputs for prediction algorithms—arises in many contexts including loan approvals, job-hiring decisions, law-enforcement-related decisions, etc. (Pasquale 2015).

Algorithms that ignore bias in inputs may provide predictions that suffer from limitations akin to those provided by human beings. Furthermore, an algorithm that ignores the biases in the input data could also exacerbate the errors stemming from those biases (Coiera 2015, Barocas and Selbst 2016). Therefore, designing an optimal decision-support algorithm that accounts for biases in the input data is critical. This approach is also consistent with the ideas expressed by Goes (2013, p. v) in a recent editorial summarizing some of the well-known cognitive biases and calling for "combining IS [information systems] research methods (explanatory, predictive, experimentation) with behavioral economics principles to shed light on the issues and inform design science research." Consistent with this theme, in this paper, we examine the design and value of an algorithm that accounts for bias in the input data. Specifically, we develop a linear classifier that accounts for the human-generated bias in the input data. We contextualize the problem in healthcare and examine the design of a bias-aware algorithm for use in a clinical decision support system (CDSS). The CDSS is a decision aid to be used by referring physicians in mammography-based breast cancer diagnosis.

1.1. Role of Bias in Breast Cancer Diagnosis

In the breast-cancer-diagnosis context, typically, referring physicians make diagnostic decisions and recommendations based on radiologists' assessments, patients' clinical-risk information, and such factors as patient preferences regarding treatment options (Ayvaci et al. 2018). The radiologist's primary task is to provide a probabilistic risk assessment based on mammogram findings. In practice, the radiologist often uses the patient's clinical-risk information, in addition to the image.[1] The medical community has debated the impact of clinical-risk information on radiologists' assessments (Loy and Irwig 2004). Some have argued that it improves the accuracy of image interpretation, but others have pointed out that it may bias the radiologist (Elmore et al. 1997). For example, Sibbald and Cavalcanti (2011) reported that the patient's clinical history may cause the radiologist's assessment to deviate from a true assessment of the mammogram.[2] The following quote from renowned radiologist Elizabeth Burnside, professor of radiology and chief of breast imaging at the University of Wisconsin–Madison, who is subspecialty trained in breast imaging, acknowledges the radiologists' bias in the breast cancer context:

> There is controversy as to whether radiologist should be weighing the clinical-risk information. First, it may be biasing them; however, it may also be helping them make correct decisions. Second, it is probable that the healthcare system, in general, is double, triple, or quadruple counting demographic risk factors as each physician factors that information into their decision-making.

In effect, the undue influences resulting from radiologists' cognitive biases may affect the performance of the referring physician who uses the radiologist's assessment. Thus, a central issue faced by the referring physician is how to combine the radiologist's assessment with the patient's clinical-risk information to make a patient-specific recommendation.

Clearly, not providing the clinical-risk information to the radiologist can potentially eliminate the bias, but this strategy will also eliminate the benefits such information provides to radiologists when interpreting mammograms. The referring physician could be made aware of the biased radiologist assessment so that she could account for the bias in her decision making. However, the referring physician is likely to be prone to her own biases; for instance, the reputation of the radiologist could sway the referring physician. Moreover, the referring physician typically deals with multiple radiologists, and accounting for the diverse characteristics of radiologists is likely to be challenging. A CDSS developed specifically to help the referring physician make a recommendation to the patient on whether a follow-up action is necessary can be a valuable tool. An important component of such a CDSS is the classification algorithm that uses the clinical-risk information and mammogram assessment to predict the likelihood of cancer or even suggest a follow-up action. The CDSS tool is also a step in the direction of the recent movement within the medical community toward widespread use of decision aids such as breast cancer risk calculators for reasons other than bias (Freedman et al. 2005, Pace and Keating 2014).

1.2. Research Questions and Contributions

A CDSS includes many components, ranging from the classification algorithm to the knowledge base and to its interactions with other applications in the host environment (Greenes 2014). However, in this research, we focus only on the classification task. In particular, we develop a linear classification algorithm[3] that uses two information sources—the radiologist's mammogram assessment and the clinical-risk information, where the mammogram assessment could be biased by the clinical-risk information—and determines whether

a patient has cancer.[4] We seek to answer three key research questions:

(i) What is the optimal design of a linear classification algorithm in the presence of radiologist bias?

(ii) How does the bias affect the algorithm's performance and design?

(iii) Should the clinical-risk information be used at all in the diagnostic process given its potential to bias the radiologist?

We answer the previous questions by first developing and analyzing a theoretical model of the breast-cancer-diagnosis context. We then quantify the theoretical results using real-life data widely used by the breast cancer medical community.

Our contributions in this research are twofold. First, our approach to designing a bias-aware classification algorithm and our analysis of the impact of bias on the algorithm and its performance are new to the literature on classification algorithms. Hajian et al. (2016) classify the adjustments to the prediction process in the presence of input noise into three categories: preprocessing of data, adjustments to the algorithm, and postprediction processing of outputs. Our approach belongs to the second category concerning adjustments to the algorithm. The information systems research literature is rich with designs and evaluations of classification algorithms for decision support (Bai et al. 2008, Gartner et al. 2015, Kao and Tang 2013, Adomavicius and Zhang 2016). In particular, our work is related to those that have examined classification under input noise (Jiang et al. 2005) and under input data distortion by strategic agents (Dalvi et al. 2004, Boylu et al. 2010, Zhang et al. 2014). However, as we further explain in Section 2.2, our approach contributes to this literature by (i) explicitly modeling the source and mechanics of noise (introduced by human biases) and (ii) separating the noise into systematic and random parts. Consequently, we are able to provide deep theoretical insights into how the input bias affects the optimal bias-aware algorithm design and the algorithm's performance.

Second, this research quantifies the value of bias-aware algorithms in the breast-cancer-diagnosis context. Rapid digitalization of healthcare after the passage of the Health Information Technology for Economic and Clinical Health Act of 2009 and the growing interest in the role of cognitive biases in medical decision making make healthcare an ideal application domain for bias-aware algorithms. The medical literature has extensively documented medical biases as well as their impacts on performance and strategies to mitigate these impacts (e.g., see Sibbald and Cavalcanti 2011). It also recognizes that CDSSs are useful in medical decision making (Bright et al. 2012, e.g., for a U.S.-government-sponsored review of CDSS trials, see), potentially also in mitigating potential harms of biases (e.g., Holmes-Rovner et al. 2007, Patel et al. 2015). Despite the interest in decision biases, to our knowledge, the literature has not rigorously examined the design of bias-aware algorithms and their value in mitigating the impact of bias. Our findings contribute to not only the design of a new class of breast cancer risk calculators that account for radiologists' bias, but also the value of such calculators in terms of decision accuracy and utility.

The rest of this paper has the following structure. We review the relevant literature and expand on our contributions in Section 2. We describe our model in Section 3. The theoretical results regarding the design and value of the optimal CDSS are discussed in Section 4. We quantify our results using real-life data in Section 5. Finally, we conclude the paper with a summary and possible future research directions in Section 6.

2. Related Literature

Our work relates to the literature about (i) cognitive biases in medical decision making, (ii) design of decision support systems to aid diagnostic decision making, and (iii) mathematical modeling of decision biases.

2.1. Cognitive Biases in Radiological Diagnosis

Two recent review articles provide a comprehensive survey of cognitive biases in medical diagnosis (Blumenthal-Barby and Krieger 2015, Saposnik et al. 2016). The context we study—the presence of clinical-risk information playing a critical role in breast-cancer-diagnostic performance—may reflect features of various types of biases. Instead of providing an extensive list of biases that reflect the mechanics of the cognitive flaw we model, we describe three common biases—anchoring, confirmation, and availability, which occur when synthesizing information for medical decisions (Ogdie et al. 2012), particularly those in radiology (Lee et al. 2013, Gunderman 2009).

The anchoring bias has been observed in numerous medical-diagnosis situations, including radiology (e.g., Lee et al. 2013, Alpert and Hillman 2004). Anchoring is a cognitive bias that represents the human tendency to overvalue a single piece of information (the anchor) and undervalue additional information when making judgments (Tversky and Kahneman 1974). This bias occurs when individuals interpret the additional information around the anchor. For example, in the context of diagnostic imaging (e.g., computed tomography scans and X-rays), the reading radiologist may be anchored by the clinical history when assessing the radiographic image. As a result of anchoring, the radiologist may assess a higher (lower) risk to the mammogram when the clinical-risk information is high (low), compared to what she would assess in the absence of clinical history. To mitigate the anchoring effect, it has

been suggested that radiologists look at the image first even though it is only a partial cure (Griscom 2002).

Confirmation bias occurs when the decision maker looks for confirming evidence to bolster an early assessment, rather than trying to disprove it. For example, an early judgment based on the patient's clinical-risk information may induce a radiologist to focus on mammogram features consistent with the initial impression and ignore the features that conflict with it (Wallsten 1981). Confirmation bias relates to a radiologist's interpretation of findings in support of prior conclusions, while anchoring relates to overreliance on a single piece of information, the clinical-risk information. In a way, confirmation bias can compound the anchoring effect. Indeed, it has been argued that confirmation bias may amplify the anchoring effect in diagnostic decisions (Croskerry 2003).

The presence of clinical-risk information at the time of mammogram interpretation could make the events associated with such information more salient. For example, a presence of breast cancer in the family history can induce the radiologist to recall relevant instances of disease, leading to availability bias (Eraker and Politser 1982). More formally, availability bias occurs when a person overestimates the probability of an event that comes immediately to the person's mind. In diagnostic decisions, when preliminary (and possibly weak) evidence suggests the presence of a disease, the radiologist may tend to overestimate the likelihood of that disease (Alpert and Hillman 2004). The clinical-risk information, when available at the time of imaging interpretation, may nudge the radiologist to be more suspicious of any findings on an image, which may or may not improve the performance. Mamede et al. (2010) found that medical residents were susceptible to availability bias.

2.2. Related IS Research

The existing literature on classification under noisy data focuses mainly on the development of expert systems where decision rules are learned from training data—also called inductive expert systems (Hong and Tsang 1997, Wu et al. 2003, Saar-Tsechansky and Provost 2007). Aside from directly addressing the input data quality issues to improve the expert system's performance (Provost 2005, Parssian 2006), methods used as a postinduction mechanism (e.g., Quinlan 1986) and methods applied during the induction process (e.g., Clark and Niblett 1989) have been proposed. Furthermore, to eliminate the negative effects of noise in training data, researchers developed a variety of input modification methods (e.g., see Mookerjee 2001, Jiang et al. 2005). While research in this stream did not generally address medical decision making, there has been some recent interest in developing machine-learning-based models to help manage chronic diseases (Meyer et al. 2014).

A substream of the classification literature focuses on discriminating classes under strategically manipulated data. Research in this substream showed that substantial gain could be achieved by designing a classifier that accounts for the incentives of the agents to fake their attributes (Dalvi et al. 2004). In the context of adversarial classification, researchers proposed methods for optimal classification (Boylu et al. 2010) or for approximate solutions (Boylu et al. 2007). Specific adversarial behavior such as that by strategic agents who alter or disguise their attributes to attain favorable decisions was also studied (Dekel et al. 2010). Subsequent work sought to develop a linear classifier in conjunction with an imputation mechanism for handling the missing data that are strategically hidden by agents for favorable decisions (Zhang et al. 2014).

Our work differs from the previously mentioned studies in several ways. First, much of the research on noise-handling techniques does not examine or model the source of noise except for research examining adversarial-classification contexts in which intelligent agents distort data strategically. We explicitly model the source of noise as similar to adversarial classification, yet the data distortion in our setup is due to shortcomings of human cognition rather than to strategic agents. Consequently, unlike the adversarial classification counterpart, the classification does not affect the behavior of agents in our context. Second, our model decomposes the source of noise into systematic and random parts (later referred to as the mean error and error variance caused by bias, respectively), which allows us to quantify the trade-off between the two. Finally, the explicit modeling of the source and mechanics of noise enables us to provide theoretical insights into how the input bias affects the optimal bias-aware algorithm design and the algorithm's performance.

2.3. Related Decision Analytic Work

The decision analysis and operations management community has recently started examining problems related to mathematical modeling of decision biases. For instance, in seeking the true parameter values of probability distributions based on expert opinion, an optimization-based approach was proposed to determine the weights for biased quantile judgments (Bansal et al. 2017). Focusing on overconfidence bias in the context of inventory decisions in a competitive environment, research has found that bias can be beneficial in coordinating the firms' decisions (Li et al. 2017). To alleviate overconfidence bias and the overfitting problem while aggregating forecasts, a machine-learning mechanism was developed (Grushka-Cockayne et al. 2016). A study related to this work descriptively examined the impact of biased mammogram information when it is combined with patient-profile information (Ayvaci et al. 2017). Yet, it did not focus on the optimal aggregation of information, nor did it consider the utility or the

optimal decision rule to classify a patient. In contrast, we take a prescriptive approach and study how the algorithm should be designed to optimally aggregate information and make recommendations in the presence of biased data. The algorithm accounts for both the mean error and error variance due to bias, and provides a recommendation to the referring physician.

3. Model Description

Consider a representative patient with an unknown true health status with regard to breast cancer. A radiologist is tasked with providing a probabilistic assessment about the presence or absence of cancer by examining the patient's mammogram. The radiologist's assessment is a quantitative measure of the risk (or likelihood) that the patient has cancer, and we refer to her assessment as the *mammogram risk*.[5] The radiologist has access to the patient's *clinical-risk information*. This is also a quantitative risk measure, but it is based solely on the patient's risk factors such as age, family history, previous breast procedures, hormone therapy, etc. The radiologist's probabilistic assessment could be biased or distorted by the clinical-risk information. Table 3 in Online Appendix A summarizes the model notation we use in this manuscript, which we describe next.

We indicate the presence or absence of cancer using labels + and −, respectively. We denote the true (unbiased) risk implied by the patient's mammogram (imaging) and the true risk implied by the patient's clinical-risk information as x_i and x_c, respectively. Conditional on the true health status of the patient, x_i and x_c follow a bivariate-normal distribution. Using $\mathcal{N}(\mu_{\cdot\cdot}, \sigma_{\cdot\cdot})$ to indicate a normal distribution with mean $\mu_{\cdot\cdot}$ and standard deviation $\sigma_{\cdot\cdot}$, we have $x_i|- \sim \mathcal{N}(\mu_{ni}, \sigma_{ni})$, $x_c|- \sim \mathcal{N}(\mu_{nc}, \sigma_{nc})$, $x_i|+ \sim \mathcal{N}(\mu_{ci}, \sigma_{ci})$, and $x_c|+ \sim \mathcal{N}(\mu_{cc}, \sigma_{cc})$. The correlation coefficient between $x_i|-$ and $x_c|-$ is $\rho_n \geq 0$, and that between $x_i|+$ and $x_c|+$ is $\rho_c \geq 0$. Without loss of generality, we assume $\mu_{ci} > \mu_{ni}$ and $\mu_{cc} > \mu_{nc}$, which implies that the mean risk levels, both mammogram and clinical, are higher for sick patients (+) than for the healthy patients (−).

When the radiologist is unbiased, her estimate of the mammogram risk for a patient will be the true x_i associated with that mammogram. Under bias, her estimate, $\hat{x}_i$, will deviate from x_i. The bias can affect the radiologist to varying degrees depending on situational factors such as the presence of a medical trainee (Hawley et al. 2016) or the radiologist's mood at the time of mammogram interpretation (Englich and Soder 2009). In support, previous work reported a notable variation in two readings by the same radiologist, referred to as the intraobserver variation (Berg et al. 2000). Hence, there could be variation or noise in the level of distortion in the radiologist's risk estimate for the same patient and the same mammogram. Consequently, the error introduced by bias is a random variable that follows a probability distribution.

Let the probability of the presence or absence of the disease in the patient population be denoted by $P(+)$ or $P(-)$, respectively. Then, the mean of clinical-risk information in the population is $\bar{x}_c = P(-)\mu_{nc} + P(+)\mu_{cc}$. Denoting the error introduced by bias in the radiologist's assessment as $\hat{x}_i - x_i$, we model the error conditional on clinical-risk information as $\hat{x}_i - x_i|x_c \sim \mathcal{N}(\beta(x_c - \bar{x}_c), \beta\sigma_0)$. The mean error, $\beta(x_c - \bar{x}_c)$, depends on two factors: (i) how much x_c deviates from its population mean, $\bar{x}_c$, which serves as the anchor, and (ii) the bias factor $\beta \geq 0$, which is a measure of the radiologist's inherent bias level. For a given bias factor, a higher-than-average (lower-than-average) value of x_c induces a positive (negative) mean error in the mammogram risk assessment, and, for a given x_c, a higher bias factor induces a higher mean absolute error. The specific features of our model, namely, the linearity in perturbations with similar slopes in the positive and the negative directions around the anchor and its continuity, are consistent with the experimental analyses of Adomavicius et al. (2013, 2014) and empirical models developed for detecting anchoring in experts' consensus forecasting in the financial context (Campbell and Sharpe 2009). The parameter σ_0 captures the variability in the error introduced by bias.[6] Specifically, $\beta = 0$ and $\sigma_0 = 0$ correspond to the unbiased assessment of x_i. Hereafter, we refer to σ_0 as *error variance* for simplicity.

We note that in the statistics literature, the mean error ($\beta(x_c - \bar{x}_c)$ in our model) is also referred to using terms such as *accuracy*, *systematic error*, and *statistical bias*. Similarly, the error variance (σ_0^2 in our model) is also referred to as *precision* and *random error*. We reserve the term *bias* to denote the radiologists' cognitive bias in this paper. We note that in our model of radiologists' bias, $\hat{x}_i$ and x_i have the same expected value of $P(-)\mu_{ni} + P(+)\mu_{ci}$ in the population; that is, while the bias distorts the mammogram risk at the individual patient level, it does not introduce any statistical bias to mammogram risk at the population level.

Given a biased assessment from a radiologist, the tasks of the classification algorithm (hereafter, *bias-aware algorithm*) are to (i) aggregate $\hat{x}_i$ and x_c and (ii) determine a threshold k for the aggregated risk to classify the instance as + or −. Following the convex linear aggregation method introduced by Stone (1961) to aggregate expert opinions, we define the aggregate information, r, as

$$r := (1 - \alpha)\hat{x}_i + \alpha x_c, \tag{1}$$

where α denotes the weight assigned to x_c, and $1 - \alpha$ denotes the weight assigned to $\hat{x}_i$.[7] An instance is classified as positive if r for that instance is above the threshold k.

Based on the definition in (1), r is normally distributed. We let $(r|-) \sim \mathcal{N}(\mu_n, \sigma_n)$ and $(r|+) \sim \mathcal{N}(\mu_c, \sigma_c)$. Then, we have

$$\left.\begin{aligned}
\mu_n &= (1-\alpha)\mu_{ni} + \alpha\mu_{nc} + (1-\alpha)\beta P(+)(\mu_{nc} - \mu_{cc}),\\
\sigma_n^2 &= (1-\alpha)^2\sigma_{ni}^2 + (\alpha + (1-\alpha)\beta)^2\sigma_{nc}^2 + (1-\alpha)^2\sigma_0^2\\
&\quad + 2(1-\alpha)(\alpha + (1-\alpha)\beta)\rho_n\sigma_{ni}\sigma_{nc},\\
\mu_c &= (1-\alpha)\mu_{ci} + \alpha\mu_{cc} + (1-\alpha)\beta P(-)(\mu_{cc} - \mu_{nc}),\\
\sigma_c^2 &= (1-\alpha)^2\sigma_{ci}^2 + (\alpha + (1-\alpha)\beta)^2\sigma_{cc}^2 + (1-\alpha)^2\sigma_0^2\\
&\quad + 2(1-\alpha)(\alpha + (1-\alpha)\beta)\rho_c\sigma_{ci}\sigma_{cc}.
\end{aligned}\right\} \quad (2)$$

In this study, our focus is on exploring the role of bias in the design and performance of the bias-aware algorithm and not the ability of either type of risk information in discriminating the + cases from the − cases. One factor that contributes to the discriminative ability of the risk information is the standard deviation of that information for the + or − cases. For brevity in mathematical notation, we assume $\sigma_i^2 := \sigma_{ni}^2 = \sigma_{ci}^2$ and $\sigma_c^2 := \sigma_{nc}^2 = \sigma_{cc}^2$, and $\rho := \rho_n = \rho_c \geq 0$.[8] Under these assumptions, we observe that $\sigma_n = \sigma_c$ holds based on the equations in (2), and we denote this quantity by σ.

For a given r and threshold, k, we let $P(\text{FN})$ denote the probability of a false negative classification and $P(\text{FP})$ denote the probability of a false positive classification. These probabilities can be calculated using

$$P(\text{FN}) = P(r \leq k \mid +)P(+) = \Phi\left(\frac{k-\mu_c}{\sigma_c}\right)P(+), \quad (3)$$

$$P(\text{FP}) = P(r > k \mid -)P(-) = \left(1 - \Phi\left(\frac{k-\mu_n}{\sigma_c}\right)\right)P(-), \quad (4)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal random variable. From the axioms of probability, we can easily verify that $P(\text{TN}) = P(-) - P(\text{FP})$ and $P(\text{TP}) = P(+) - P(\text{FN})$, where $P(\text{TN})$ and $P(\text{TP})$ denote the probabilities of true negative and positive, respectively. We study two different performance objectives for the bias-aware algorithm: (i) expected utility and (ii) area under the receiver operating characteristic curve (AUC), which measures the discriminative ability.

Utility maximization is a commonly used objective in various decision contexts including business, economics, and healthcare (Clemen and Winkler 1999, Ranjan and Gneiting 2010, Kardaras and Žitković 2011). To maximize the utility, one needs to optimize over α and k. We denote the utility for possible classification outcomes using U_{TP} (true positive), U_{FP} (false positive), U_{TN} (true negative), and U_{FN} (false negative). Motivated by the clinical intuition that correct diagnosis of diseased patients is valued more than correct diagnosis of healthy patients, we assume $U_{\text{FN}} < U_{\text{TP}} < U_{\text{FP}} < U_{\text{TN}} < 0$. We use negative values for utility to indicate the costs associated with outcomes. In the case of breast cancer, the most costly outcome is a false negative, where the patient is falsely predicted to be free of cancer. A false negative assessment may lead to significant loss of expected life because of the delay in the proper treatment of the patient. On the other extreme, the least costly outcome is a true negative assessment, where the patient is correctly predicted as healthy. Under a true-negative outcome, the cost incurred by the patient is the time lost to get the mammography and exposure to the radiation. Associating a disutility to true negative outcomes is typical to health-economics studies (Gold 1996). Under a false positive assessment, the patient will incur the additional costs of undergoing biopsy despite the absence of the disease, which will be larger than a true negative in absolute terms. On the other hand, when the patient is correctly diagnosed (true positive), she will be worse off than a healthy patient with incorrect diagnosis (false positive) while being better off than an untreated sick patient (false negative). The expected utility is computed to be

$$\begin{aligned}
&E[U(\alpha,\beta,\sigma_0,k)]\\
&\quad = U_{\text{TP}}P(\text{TP}) + U_{\text{FP}}P(\text{FP}) + U_{\text{FN}}P(\text{FN}) + U_{\text{TN}}P(\text{TN}),\\
&\quad = (U_{\text{TP}} - U_{\text{FN}})P(\text{TP}) + (U_{\text{FP}} - U_{\text{TN}})P(\text{FP})\\
&\qquad + U_{\text{FN}}P(+) + U_{\text{TN}}P(-). \quad (5)
\end{aligned}$$

Denoting the optimal expected utility for given β and σ_0 as $U^*(\beta,\sigma_0)$, we formally state the utility maximization problem as

$$U^*(\beta,\sigma_0) = \max_{\alpha,k} E[U(\alpha,\beta,\sigma_0,k)]. \quad (6)$$

The AUC is a well-recognized measure of discriminative ability of classification (Ulvila and Gaffney 2004, Cavusoglu et al. 2005) and is used widely in diagnostic decisions based on radiological imaging (Cantor et al. 1999). The AUC measure is independent of k (Metz 1978). Let $AUC^*(\beta,\sigma_0)$ be the optimal discriminative ability in terms of AUC for given β and σ_0. We formally state the AUC maximization problem as

$$AUC^*(\beta,\sigma_0) = \max_{\alpha} AUC(\alpha,\beta,\sigma_0), \quad (7)$$

where the AUC value of the aggregate distribution can be calculated (see Metz 1978) using

$$AUC(\alpha,\beta,\sigma_0) = \Phi\left(\frac{\mu_c - \mu_n}{\sqrt{2\sigma^2}}\right). \quad (8)$$

Since $\Phi(\cdot)$ is strictly increasing, maximizing the function $AUC(\alpha,\beta,\sigma_0)$ is equivalent to maximizing $(\mu_c - \mu_n)/\sqrt{2\sigma^2}$.

4. Theoretical Analysis

We first examine the optimal bias-aware algorithm based on the AUC and the expected utility objectives. We then examine how bias affects algorithm performance

and quantify the impact of ignoring bias while designing the algorithm. Next, we identify the conditions under which it is suboptimal to use the clinical-risk information given that it can bias the radiologist's assessment of the mammogram risk. Finally, we discuss the case when bias impacts error mean and error variance proportionally.

For clarity in exposition, we define some additional notations and provide intuitive interpretations for them. For any α, β, and σ_0, we define $I(\alpha,\beta,\sigma_0) = (\mu_c - \mu_n)/\sigma$ as the discriminative ability of r because, as we will show later, the quantity $I(\alpha,\beta,\sigma_0)$ is proportional to the *AUC* when r is used for classification. Note that $I(0,0,0)$ corresponds to the discriminative ability of only mammogram risk, whereas $I(1,0,0)$ corresponds to that of only clinical-risk information. We define the differences in the mean risks for the healthy and sick populations due to mammogram and clinical-risk information using $\Delta_i = \mu_{ci} - \mu_{ni}$ and $\Delta_c = \mu_{cc} - \mu_{nc}$, respectively, and let

$$h := \frac{I(0,0,0)}{I(1,0,0)} = \frac{\Delta_i}{\sigma_i}\Big/\frac{\Delta_c}{\sigma_c} \quad \text{and}$$
$$t := \frac{(U_{TP}-U_{FN})/P(-)}{(U_{TN}-U_{FP})/P(+)} = \frac{U_{TP}-U_{FN}}{U_{TN}-U_{FP}}\cdot\frac{P(+)}{P(-)}. \tag{9}$$

The quantity h denotes the discriminative ability of the mammogram risk relative to that of the clinical-risk information, and t represents the relative value (utility) of detecting the disease (+ cases) compared to reassuring its absence (− cases).

4.1. Optimal Bias-Aware Algorithm

Let $k(\alpha,\beta,\sigma_0)$ be the optimal threshold such that when r exceeds k, the algorithm will classify the instance as + for given relative weight for the clinical-risk information α, bias factor β, and error variance σ_0. Lemma 1 in Online Appendix C shows that

$$k(\alpha,\beta,\sigma_0) := \frac{\mu_n+\mu_c}{2} - \frac{\ln(t)\sigma^2}{\mu_c-\mu_n} \tag{10}$$

holds. The following theorem characterizes the optimal weight and threshold.

Theorem 1. *Define $\alpha^*(\beta,\sigma_0)$ and $k^*(\beta,\sigma_0)$ as*

$$\alpha^*(\beta,\sigma_0) := \frac{1}{1+\dfrac{\sigma_i\sigma_c(h-\rho)}{\sigma_i^2(1-h\rho)+\sigma_0^2-\sigma_i\sigma_c\beta(h-\rho)}}, \tag{11}$$
$$k^*(\beta,\sigma_0) := k(\alpha^*(\beta,\sigma_0),\beta,\sigma_0). \tag{12}$$

Then, $\alpha^*(\beta,\sigma_0)$ and $k^*(\beta,\sigma_0)$ maximize the total expected utility, and $\alpha^*(\beta,\sigma_0)$ maximizes the AUC.

Theorem 1 provides (i) the optimal weight $\alpha^*(\beta,\sigma_0)$ that should be assigned to the clinical-risk information and (ii) the optimal aggregate risk threshold $k^*(\beta,\sigma_0)$ for classifying the patient as +. We have four observations from Theorem 1:

(i) The utility parameters affect only the optimal risk threshold and not the optimal aggregation weight, implying that the utility parameters are required only for classifying the patient into + or −, but they are not required to compute the aggregate risk score based on clinical-risk information and (biased) mammogram risk. This result shows that the algorithm can use a staggered approach, determining first the optimal aggregation to maximize the discriminative ability of the aggregate risk followed by choosing the optimal risk threshold to maximize the expected utility. The result suggests that uniform guidelines can be provided to generic risk calculators regarding how to aggregate the clinical-risk information and the mammogram risk regardless of the specific patient utility characteristics. However, the risk threshold will have to be tailored for each patient depending on her specific utility parameters.

(ii) If $t = 1$, that is, if the value from correctly classifying + cases is equal to that from − cases, then the risk threshold is also independent of the utility parameters.

(iii) When the radiologist is unbiased such that $\beta = 0$ and $\sigma_0 = 0$, the optimal weight for the clinical-risk information reduces to that shown by Winkler (1981) when examining Bayesian aggregation of expert information under no bias.

(iv) After some algebraic manipulation, one can observe that the optimal weight for the clinical-risk information can be negative depending on the relative discriminatory ability of the two risks (h), the correlation between them (ρ), and the bias parameters (β and σ_0). An increase in the correlation between the two risks suggests that the mammogram risk captures more of the information content of the clinical-risk information. Therefore, the clinical-risk information provides less incremental value with such an increase in correlation. In contrast, when the two risks are highly correlated, the use of clinical-risk information compounds, rather than compensates for, the bias-induced error in the mammogram risk in classifying patients. In such cases, a negative weight to the clinical-risk information becomes optimal.

Even in the absence of bias, the optimal weight assigned to clinical-risk information is negative when $\rho > \min(h, 1/h)$. Because clinical-risk information and mammogram risk play complementary roles in breast cancer diagnosis, we assume $\rho < \min(h, 1/h)$ in the rest of our theoretical analysis to ensure that both of these information types are assigned a positive weight while classifying the patient in the absence of bias. Clearly, if the two risks are conditionally independent, which is assumed widely in the classification (see, e.g., Jiang et al. 2005) and medical (see, e.g., Burnside et al. 2009) literatures, such an assumption is satisfied.

Structural analysis of the optimal weight given in Equation (11) shows the following. First, the optimal weight assigned to the clinical-risk information, $\alpha^*(\beta, \sigma_0)$, is decreasing in bias factor β. An increase in the bias factor increases the influence of the clinical-risk information in the radiologist's mammogram risk assessment, and therefore, the optimal bias-aware algorithm counteracts this influence by reducing the weight assigned to the clinical-risk information in its decision-making process. Second, in contrast, an increase in the error variance (σ_0) increases the weight assigned to the clinical-risk information. This is because, under bias, an increase in σ_0 increases the variability of the (biased) mammogram risk, which, in turn, reduces the relative precision of the mammogram risk compared to the clinical-risk information. Finally, an increase in the correlation between the clinical-risk information and the mammogram risk (ρ) decreases the weight assigned to the clinical-risk information when the error variance is not too high. Such a finding is intuitive because an increase in the correlation makes the information content of the clinical-risk information less valuable. While we can analytically characterize the impact of model parameters on the optimal aggregation weight, a theoretical analysis of the effect of model parameters on the decision threshold proved to be intractable. Numerical analysis suggests that the risk threshold for classifying a patient into the + category, $k^*(\beta, \sigma_0)$, increases when either bias factor β or error variance σ_0 increases.

4.2. Impact of Bias Under the Optimal Bias-Aware Algorithm

Proposition 1 characterizes how the bias factor, error variance, and other model parameters affect the expected utility and the discriminative ability under the optimal bias-aware algorithm.

Proposition 1. *Let $\beta_1 > \beta_2 > 0$ and $\sigma_0 > 0$. Then,*

(i) $U^*(\beta_2, \sigma_0) = U^*(\beta_1, \sigma_0) < U^*(\beta_1, 0) = U^*(0, 0)$,

(ii) $AUC^*(\beta_2, \sigma_0) = AUC^*(\beta_1, \sigma_0) < AUC^*(\beta_1, 0) = AUC^*(0, 0)$,

(iii) $U^*(\beta, \sigma_0)$ *is independent of* β *and decreases in* σ_0, σ_i, σ_c, *and* ρ.

According to Proposition 1, (i) and (ii), the expected utility and discriminative ability are (weakly) smaller in the presence of bias than in its absence even under the bias-aware algorithm. This result demonstrates that the radiologist's bias can only diminish the performance, despite accounting for the bias using the algorithm. Proposition 1, (i) and (ii), further shows that when error variance is zero (i.e., $\sigma_0 = 0$), the optimal bias-aware algorithm eliminates the negative impact of bias. On the other hand, in the presence of a positive error variance (i.e., $\sigma_0 > 0$), the negative effect of bias cannot be eliminated through the optimal bias-aware algorithm alone. This result suggests that if the radiologist providing the imaging risk assessment has highly predictable behavior, then, instead of trying to eliminate the bias at the radiologist's end (which could be impossible or costly), a CDSS can support the referring physician by employing the optimal bias-aware algorithm.

We find from Proposition 1(iii) that the optimal expected utility, $U^*(\beta, \sigma_0)$, is independent of the bias factor (and, therefore, the mean error), yet it decreases in the error variance. The result reinforces the implication that bias-reducing efforts are useful even when the algorithm (therefore the CDSS) accounts for the bias. Furthermore, we find that the optimal expected utility decreases in σ_i, σ_c, and ρ. The findings are intuitive because an increase in the standard deviation of either risk or the correlation between the two risks reduces the information content of the aggregated risk.

4.3. Impact of Ignoring Bias in the Algorithm

We define the value of the bias-aware algorithm as the optimal utility when the algorithm adjusts for bias less the utility when the algorithm assumes that bias does not exist (i.e., the algorithm that uses $\alpha^*(0, 0)$ and $k(\alpha^*(0, 0), 0, 0)$ despite nonzero β and σ_0). We refer to the latter as the *bias-blind* algorithm. We formally define the utility value of the bias-aware algorithm (compared to the bias-blind algorithm) as

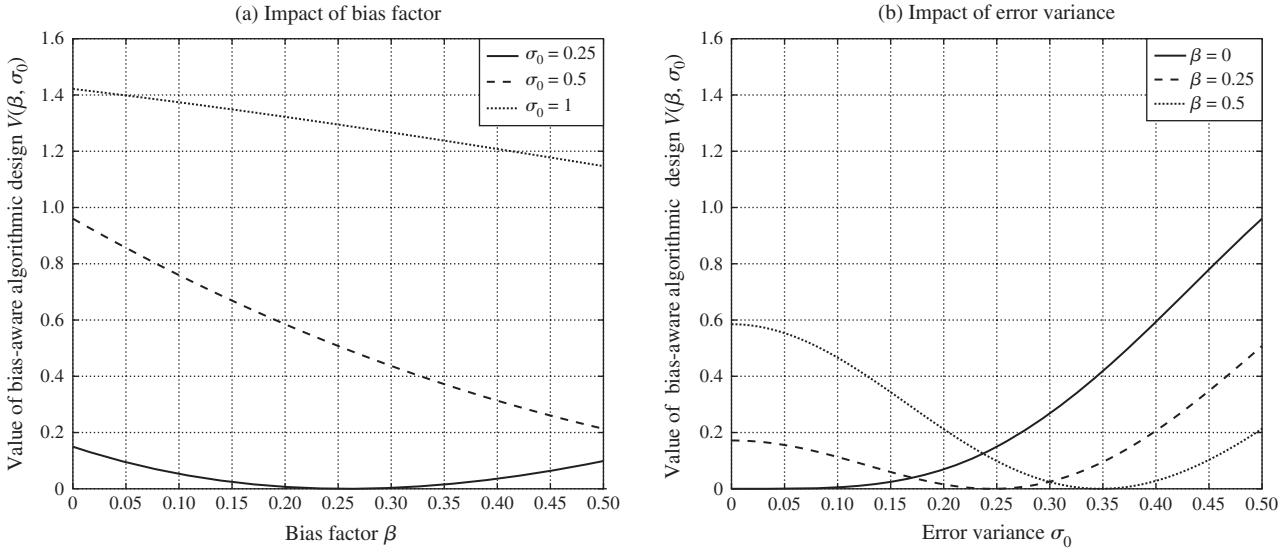
$$V(\beta, \sigma_0) := U^*(\beta, \sigma_0) - U(\alpha^*(0, 0), \beta, \sigma_0, k(\alpha^*(0, 0), 0, 0)). \tag{13}$$

The next proposition characterizes how the bias affects the value of the bias-aware algorithm.

Proposition 2. *The bias-aware algorithm has* (i) $V(\beta, \sigma_0) \geq 0$, *and* (ii) $V(\beta, \sigma_0)$ *is decreasing in* β *when* $\beta < \sigma_0^2/(\sigma_c \sigma_i \cdot (h - \rho))$ *and increasing otherwise.*

Proposition 2 suggests that the optimal bias-aware algorithm offers a nonnegative value and that there is a bias factor threshold below which the value decreases in the level of bias factor and increases otherwise. A noteworthy observation from the finding is that the value does not necessarily increase in the bias factor (β). If there is no error variance ($\sigma_0 = 0$), then an increase in mean error always increases the value of the bias-aware algorithm. On the other hand, when the error variance is not zero, the value increases in the mean error only when the mean error is larger than a threshold. The intuition for this result is our earlier finding (Proposition 1): the adverse effect of mean error can be completely eliminated by a bias-aware algorithm; namely, $U^*(\beta, \sigma_0)$ is independent of β where the bias-aware algorithm adjusts α and k to achieve the same $U^*(\beta, \sigma_0)$ when β is changed. However, an increase in the mean error increases the performance difference between the bias-aware and bias-blind algorithm only when the error variance is not too high—essentially, an increase in mean error mitigates

Figure 1. The Value of Accounting for Bias as a Function of Bias Factor (β) and Error Variance (σ_0)



Note. For the figure, we use the parameters estimated in Section 5 when $AUC_c = 0.656$ and $AUC_i = 0.820$.

the adverse impact of a large error variance for fixed values of α and k.

Proposition 2 shows that the interaction between the mean error and error variance due to bias is critical in determining the value of the bias-aware algorithm. Figure 1, (a) and (b), provides a numerical example of the value as a function of bias factor and error variance, respectively. The graphs illustrate the theoretical finding in Proposition 2 and also provide insights about the impact of error variance on the value of the bias-aware algorithm, which we could not theoretically analyze. It suggests that, generally, the value is high when the mean error is low and error variance is high, and it is low when the mean error is moderate for a given error variance.

4.4. Should the Clinical-Risk Information Be Used Under Radiologist Bias?

The previous analysis showed that the adverse impact of mean error alone due to bias can be eliminated, but an increase in the error variance diminishes the usefulness of the clinical-risk information. Thus, even though the clinical-risk information may be useful on its own, it is unclear whether the clinical-risk information should be used when it has the potential to bias the radiologists' risk assessment of a mammogram. A social planner, designing the whole system, can consider two possible alternatives: (i) use the clinical-risk information but design the algorithm to account for the bias, and (ii) not use the clinical-risk information at all (neither the radiologist nor the algorithm uses it) when the social planner cannot prevent parts of the healthcare system from using the clinical-risk information. Clearly, the option in which the clinical-risk information is withheld from the radiologist while the algorithm uses it will lead to the maximum possible utility. However, this scenario is not practical, especially for in-house radiology situations. At the time of reading an image, the radiology information systems generally have the clinical-risk information readily available to the radiologist (Nance et al. 2013). In addition, the radiologist may choose to access to the electronic health record (EHR) directly, which occurs in as many as 73% of the imaging interpretations (Lin et al. 2010). A survey of radiologists suggests that most of them usually or always use clinical-risk information when interpreting mammograms (Carney et al. 2012). Thus, preventing radiologists from accessing the clinical-risk information when the relevant information already resides in the EHR—and is, therefore, available to an algorithm—is not likely to be a feasible option.

Despite the aforementioned concerns, assessing the value, or lack thereof, of using clinical-risk information when it biases decisions is still useful because of following. First, such an assessment provides a benchmark for identifying conditions under which clinical-risk information creates net value (benefits minus bias-induced harms being nonnegative). Second, it contributes to the understanding of broader settings where bias-prone inputs are used by algorithms and the elimination of a bias-inducing information source is a feasible option.

The case in which the system does not use the clinical-risk information is captured in our model by setting $\alpha = \beta = \sigma_0 = 0$. The following proposition characterizes when the clinical-risk information should be available in the system despite the possibility that it may bias the radiologist.

Proposition 3. *If* $h \geq 1$ *and*

$$\frac{1}{h} - \sqrt{\frac{(h^2-1)\sigma_0^2}{\sigma_i^2}} < \rho,$$

then $U^*(\beta, \sigma_0) < U(0,0,0,k(\alpha^*(0,0),0,0))$; *otherwise,* $U^*(\beta, \sigma_0) > U(0,0,0,k(\alpha^*(0,0),0,0))$.

Proposition 3 shows that when (i) the relative information content of mammogram is larger than that of the clinical-risk information and (ii) the correlation between the two information sources is large enough, the expected optimal utility under bias-aware algorithm will be less than the same when clinical-risk information is not used. The proposition offers several insights. First, it reveals that even when the clinical-risk information is informative on its own, not using it in the whole process may be better than using it. Clearly, the adverse effect arising from bias may outweigh the benefits of using the clinical-risk information. Second, as per the insights from previous analyses, a high-enough correlation will cause the clinical-risk information to be not useful. An increase in the error variance (σ_0) enlarges the region in the parameter space where not using the clinical-risk information is better than using it. Finally, the literature on induction under noise identifies an information source's discriminative ability and acquisition cost as important factors to be considered when deciding to use it in a classification task such as radiological diagnosis (see, e.g., Sheng et al. 2008). Proposition 3 shows that when a diagnostic task is based on data derived from human experts susceptible to cognitive biases, the error variance induced by bias is a key factor to consider while deciding to use an information source in diagnostic systems.

4.5. Joint Influence of Bias on Both the Mean Error and Error Variance

The baseline model introduced in Section 3 assumes the radiologist's bias influences the mean error and error variance in mammogram assessment independently. However, since the source of the error is the same agent, the radiologist in this case, a more biased radiologist is likely to influence both mean and variance similarly. Essentially, larger (smaller) mean errors will likely have larger (smaller) error variances, corresponding to a fixed coefficient of variation. In the context of errors in expert judgment, several prior studies assume a fixed coefficient of variation. For example, MacCormack and Verganti (2003) and Fisher and Raman (1996) assume proportionality of the standard deviation of forecast errors to their mean. Following this line of work, we assume that a shift in the bias factor has a proportional impact on mean error and error variance. We therefore adjust the modeling of error due to bias conditional on clinical-risk information as $\hat{x}_i - x_i | x_c \sim \mathcal{N}(\beta(x_c - \bar{x}_c), \beta\sigma_0)$. This is a special case of the baseline model, and the optimal bias-aware algorithm under the special case is obtained by replacing σ_0 with $\beta\sigma_0$ in Theorem 1. Theorem 2 in Online Appendix B provides the optimal weight and risk threshold under the new model. The results under the base model carry over to the new model with very few exceptions. When mean error and error variance are decoupled, we find that the weight $\alpha^*(\beta, \sigma_0)$ is decreasing in β. With dependent mean error and error variance, $\alpha^*(\beta, \sigma_0)$ can be increasing in β when β is sufficiently small and decreasing otherwise. The following result provides the analogue of Proposition 1 under the new model.

Proposition 4. *Under the assumption of proportional mean error and error variance, for fixed values of bias factor* β_1 *and* β_2 *satisfying* $\beta_1 > \beta_2 > 0$ *and error variance* $\sigma_0 > 0$,

$$U^*(\beta_2, \sigma_0) < U^*(\beta_1, \sigma_0) < U^*(\beta_1, 0) = U^*(0,0), \quad (14)$$

$$AUC^*(\beta_2, \sigma_0) < AUC^*(\beta_1, \sigma_0) < AUC^*(\beta_1, 0) = AUC^*(0,0). \quad (15)$$

Proposition 4 suggests that when mean error and error variance are proportional, an increase in the bias factor always hurts the optimal expected utility or discriminative performance unless there is no error variance. Hence, the main additional insight from the proposition is that when the bias factor influences both the mean error and error variance, the bias-aware algorithm alone cannot completely eliminate the adverse impact of bias. Furthermore, we find $U^*(\beta, \sigma_0)$ is decreasing in bias factor β as opposed to being independent of it. Finally, Proposition 7 in Online Appendix B is an analogue to Proposition 3 and suggests that the bias factor also determines whether or not using the clinical-risk information is valuable.

5. Design and Value of the Bias-Aware Algorithm for Breast Cancer Diagnosis: A Computational Experiment

Breast cancer is the leading cause of non-skin-cancer-related deaths among women. American Cancer Society (2014) estimates that 40,000 deaths were due to breast cancer in the year 2014, and that one in eight women is expected to develop breast cancer in her lifetime. Early screening and detection using clinical-risk information and mammography is considered to be the most effective and efficient method to prevent breast-cancer-related deaths. Models such as the Gail risk model (National Cancer Institute 2014a) that provide breast cancer risk estimates based on clinical-risk information such as family history, reproductive history, menopausal status, prior breast procedures, age, ethnicity, and others (Tice et al. 2008, Engmann et al. 2017) have a discriminative ability of about 0.650 when measured using the AUC.

The risk based on findings from mammography is generally reported using a standardized tool called Breast Imaging and Reporting Data System (BI-RADS) (American College of Radiology 1998), which has a better discriminative ability than clinical-risk models. Nevertheless, mammography is also imperfect with a false positive rate of 66.2% and a false negative rate of 14.6% (Rosenberg et al. 2006). The uncertainty associated with diagnostic data in mammography-based diagnosis, combined with the growing emphasis on evidence-based medicine, has brought significant attention to the development of CDSSs to support breast cancer screening and diagnosis (Freedman et al. 2005, Ayer et al. 2010a, Chhatwal et al. 2009). Furthermore, radiologists' possible bias when interpreting mammograms makes screening and diagnosis even more challenging (Carney et al. 2012), while also making the design of a bias-aware algorithm relevant. In this section, we illustrate the impact of bias in the design and value of a bias-aware algorithm for breast cancer diagnosis using a breast-cancer-outcomes database based on clinical-risk information and the medical literature.

5.1. Parameter Estimation

We use the Breast Cancer Surveillance Consortium (BCSC) database to estimate the parameters of the clinical-risk distributions for cancer and cancer-free patients (Ballard-Barbash et al. 1997). We do not have access to patient-level radiologist assessments; hence, we impute the parameters of the mammography risk distributions by adjusting the parameters of the clinical-risk distributions to achieve an overall mammography performance reported in the medical literature.

5.1.1. Estimating Distribution Parameters. The BCSC database has information on 1,007,600 women undergoing a total of 2,392,998 screening mammographies during the period 1996–2002. The clinical-risk factors included in the database are provided on the BCSC website (http://www.bcsc-research.org). Since each screening mammogram was linked with the cancer-registry information that records any occurrence of breast cancer diagnosis in the 12 months following the first visit, we are able to identify which of the patients in the BCSC database were diagnosed with cancer. Using the Barlow et al. (2006) estimation procedure, we used logistic regression to predict the risk of developing cancer within one year based on the clinical-risk factors. We repeated this prediction for each data point in the BCSC data set to create the risk distributions for the patients with and without cancer. Following the biostatistics literature, we applied a Box–Cox type power transformation to the estimated distributions and obtained the corresponding normal distributions for the clinical-risk information (Faraggi and Reiser 2002). Our parameter estimates for the clinical-risk distributions are $\mu_{pc} = 2.4042$, $\mu_{nc} = 2.1900$, $\sigma_{pc} = 0.3656$, and $\sigma_{nc} = 0.3869$. The indicated *AUC* of the clinical-risk information (AUC_c) is 0.656, which is consistent with the reported discriminative ability of the clinical-risk information (Barlow et al. 2006).

We impute the parameters for the mammography distributions using the performance benchmarks reported in the medical literature. Specifically, we simultaneously adjusted (i) the difference between the mean risk scores for the cancer and noncancer patients and (ii) the variance for the clinical-risk distributions to achieve the AUC performance of mammography reported in practice. According to the statistics from the National Cancer Institute, the AUC performance of the mammography (AUC_i) varies depending on the type of mammography—film or digital—and the age of the patient (National Cancer Institute 2014b). In particular, we selected three different *AUC* values for the mammography risk to represent low ($AUC_i = 0.780$), moderate ($AUC_i = 0.820$), and high ($AUC_i = 0.890$) discriminative ability. The low and high mammography performance reflect the typical range reported in performance benchmark studies (Pisano et al. 2005, Ichikawa et al. 2010). The moderate mammography performance is obtained from a study where the measurement is based solely on interpretation of mammograms and nonuse of the clinical-risk information (Lo et al. 1999). We chose the final mammography distribution parameter values so that they are comparable to the parameters of the clinical-risk information and are in line with clinical intuition. As an example, the corresponding parameters for mammography risk with high discriminative ability are given as $\mu_{pi} = 2.5393$, $\mu_{ni} = 2.0607$, $\sigma_{pi} = 0.2799$, and $\sigma_{ni} = 0.2644$.

We estimated the correlation between mammogram and clinical-risk information, ρ, from the clinical literature as 0.0548 using Spearman's rank correlation. This estimate is based on the mammography assessments and the clinical-risk information for cancer and cancer-free patients where clinical-risk information use was controlled at the time of mammography interpretation (Carney et al. 2012).

5.1.2. Estimating Utility Parameters. We defined the disutility as the life years lost, following the medical literature. To estimate the disutility values for cancer patients, we used the Wisconsin Breast Cancer Epidemiology Simulation Model estimate (Fryback et al. 2006), where a 55-year-old patient was simulated to calculate the expected life years lost with a false negative (Wu et al. 2011). We then applied a treatment-effectiveness factor to obtain the life years lost when the same patient is treated (Haybittle 1998); namely, we use $U_{TP} = -463$ days and $U_{FN} = -927$ days. Based on the clinical literature, we use $U_{TN} = -0.25$ days (Mandelblatt et al. 1992) and $U_{FP} = -7$ days (Gram et al. 1990).

5.1.3. Estimating Bias Parameters. We estimated the bias factor based on experimental results reported by Elmore et al. (1997). The experiment was specifically designed to determine whether radiologists are biased by patients' clinical-risk information when interpreting mammograms. We identified $\beta = 0.1$ as a small bias factor and $\beta = 0.25$ as a large bias factor based on the results of this experiment. While variability in radiologists' biased mammogram estimates has been reported in the literature (Carney et al. 2012), there is no quantitative estimate of it available. We use σ_0 values of 0, 0.25, 0.5, and 1 to represent no, low, moderate, and high error variance due to bias in our numerical experiments. We adopt the model where both mean error and error variance are influenced by the bias factor, as discussed in Section 4.5, for our computational experiments.

5.2. Optimal Bias-Aware Algorithm: Aggregation Weights and Decision Threshold

Our model suggests that for the prevailing clinical-risk model with an approximate *AUC* of 0.656 and an average-quality mammogram with an *AUC* of about 0.820, the clinical-risk information should carry a relative weight of approximately 24% and the mammogram risk should carry a relative weight of 76% under no bias. Furthermore, under no bias, a follow-up should be recommended for a patient if the weighted risk score (obtained using the previously mentioned relative weights) exceeds a value of 2.62. Figure 2 illustrates how the presence of bias affects the optimal bias-aware algorithm. It shows that the optimal weight on the clinical-risk information (mammogram risk) generally decreases (increases) when either bias factor (β) increases or error variance (σ_0^2) decreases. However, interestingly, our computational experiments also show that when both bias factor and error variance are sufficiently high, it is optimal to use a larger weight for the clinical-risk information under bias than under no bias. Regardless of bias parameters, we find that the threshold level for risk to recommend a follow-up is increasing in both bias factor and error variance.

A key practical implication of the computational results shown in Figure 2 is that an accurate estimation of bias parameters is essential for the optimal design of a bias-aware algorithm in the breast cancer context because the qualitative nature of the impact of bias on the design is not uniform across all bias parameters; namely, depending on the parameter values, bias can either increase or decrease the optimal weights that should be assigned to the two risk scores. The conventional view is that when the radiologist's mammogram assessment is influenced by the clinical-risk information, the risk assessment based on the mammogram already partially incorporates the clinical-risk information. Therefore, the clinical-risk information (mammogram risk) should carry less (more) emphasis when the referring physician aggregates the two sources of information to mitigate "double counting" of clinical-risk information. However, this conventional view is not valid when the bias is sufficiently high with highly variant bias-induced error; we find, in fact, that more (less) emphasis should be placed on clinical (mammogram) risk in this case. Essentially, a high bias-induced mean error and error variance[9] makes the mammogram risk so unreliable that it is better to rely more on the clinical-risk information. While the bias parameters also influence the threshold risk score for a positive assessment, unlike the relative weights on clinical-risk information and mammogram risk, an increase in any bias parameter only increases the threshold.

Figure 2. Optimal Bias-Aware Algorithm Parameters as a Function of Bias Factor for Various Levels of Error Variance ($AUC_c = 0.656$, $AUC_i = 0.820$)

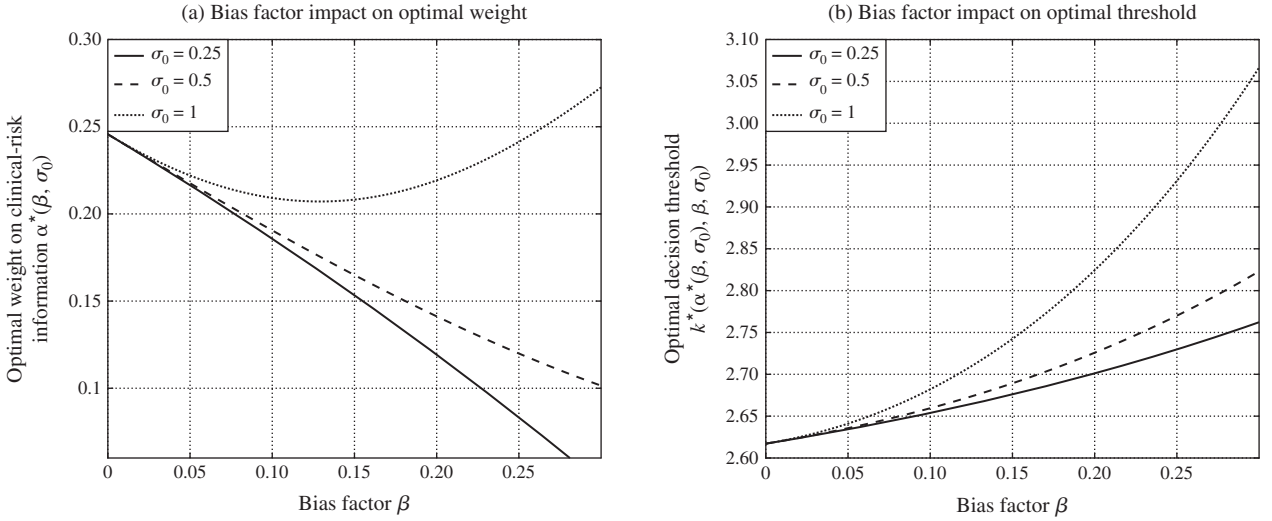


Table 1. Expected Life Years Lost for a Population of One Million Because of Radiologist Bias ($AUC_c = 0.656$)

	Patient population	$AUC_i = 0.780$	$AUC_i = 0.820$	$AUC_i = 0.890$
$\beta = 0.1$	All patients	4.63	8.00	16.69
$\sigma_0 = 0.5$	Cancer patients	7.88	11.28	17.42
$\beta = 0.1$	All patients	17.19	29.66	61.83
$\sigma_0 = 1.0$	Cancer patients	29.78	42.63	65.95
$\beta = 0.25$	All patients	70.08	119.69	249.44
$\sigma_0 = 1.0$	Cancer patients	132.72	190.20	297.75

5.3. Impact of Radiologists' Bias on Breast-Cancer-Diagnosis Outcomes

We quantify the impact of radiologists' bias as the reduction in expected life years because of the presence of bias, that is, expected life years in the absence of bias minus expected life years in the presence of bias. We assume that we use the respective optimal bias-aware algorithm in both the no-bias scenario and the bias scenario. Mathematically, this corresponds to the quantification of $U^*(0,0) - U^*(\beta,\sigma_0)$. Table 1 provides the reduction in total patient life years attributable to bias for a hypothetical mammography screening population of one million patients with 128 new cases of breast cancer per 100,000 patients (Siegel et al. 2014).[10] Table 1 reveals that when the accuracy of the mammography is moderate (i.e., $AUC_i = 0.820$), for the prevailing clinical-risk models with $AUC_c = 0.656$, the presence of various levels of bias could result in a reduction in the expected life years for all patients ranging from 8.00 to 119.69. For cancer patients alone, the range is 11.08–190.20. The values suggest that the error introduced by bias—even when accounted for—harms cancer patients as a result of cancers missed because of false negatives, while it saves the healthy patients from unnecessary follow-ups because of a reduction of false positives. The adverse impact of bias increases as the mean error or error variance due to bias increases. Furthermore, we find that the negative impact of bias is higher if the mammogram is more accurate. For instance, when $AUC_i = 0.890$, a high bias could result in a reduction of 249.44 life years for the whole screening population, and 297.75 life years for if the population consisted of only cancer patients.

Bias has a similar impact on diagnostic accuracy. When $AUC_c = 0.656$ and $AUC_i = 0.820$, the overall *AUC* is 0.837 if there is no bias. However, the overall *AUC* reduces to 0.831 under a large bias (i.e., $\beta = 0.25$ and $\sigma_0 = 1$). Although the absolute reduction in *AUC* because of bias may seem small, a back-of-the-envelope calculation shows that this level of impact on *AUC* is significant. For a screening population of about 39 million patients going through mammography in a given year (U.S. Food and Drug Administration 2015), the reduction in *AUC* from 0.837 to 0.831 translates into an additional 237,900 misdiagnoses of patients' health.

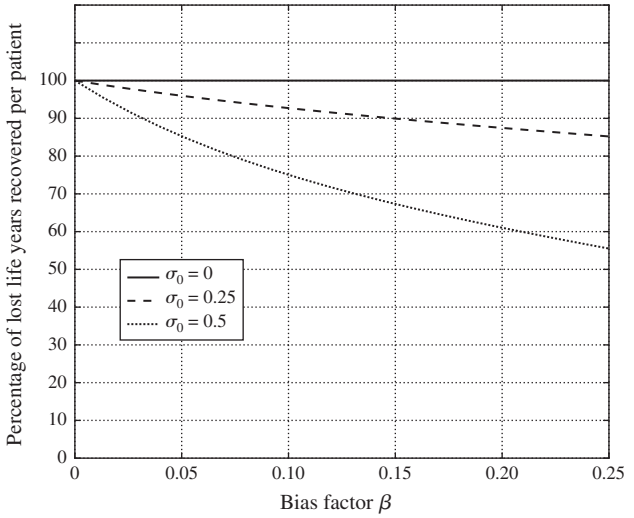
The magnitude of impacts of bias indicates that the adverse effects of radiologists' bias are substantial in terms of life years lost. However, estimating the financial impacts of bias is challenging as there is no consensus on how to value life years. Despite this, the results suggest that appropriate mechanisms to eliminate or mitigate the bias should be adopted. We next examine the value of designing a bias-aware algorithm in terms of the extent to which it mitigates the adverse impact of bias.

5.4. Value of a Bias-Aware Algorithm for Breast Cancer Diagnosis

We examine the value of a bias-aware algorithm for breast cancer diagnosis using the percentage of lost expected life years recovered by accounting for bias. We note that the no-bias case ($\beta = 0$ and $\sigma_0 = 0$) results in the maximal expected life years. Let L be the reduction in expected life years because of bias; that is, L is the expected life years under no bias minus the same under bias when the algorithm does not adjust for it (earlier referred to as the bias-blind algorithm). The bias-aware algorithm that optimally accounts for bias then recovers some of this loss. Let R be the lost expected life years that can be recovered from L if the algorithm accounts for the bias optimally. Clearly, $R \leq L$. We denote by R/L the *percentage of lost life years recovered per patient*, which represents the benefit provided by the bias-aware algorithm. We show how the percentage of lost life years recovered per patient changes with respect to mean and variance of error due to bias in Figure 3, where the *AUC* of the clinical-risk information is fixed at 0.656, and that of the mammography risk is fixed at 0.820.

We observe that when there is no error variance, $\sigma_0 = 0$, the bias-aware algorithm is able to recover 100% of the loss incurred by a bias-blind algorithm. This is consistent with Proposition 4, which shows that a bias-aware algorithm can achieve the maximum utility obtainable under no bias if $\sigma_0 = 0$. We also observe that the value of a bias-aware algorithm diminishes when either the mean error or error variance increases, but the value can still be substantial. The rate of reduction in the value of a bias-aware algorithm due to an increase in β is slower when σ_0 is smaller. In particular, when the mean error due to bias is small ($\beta = 0.1$), the bias-aware algorithm recovers 92.4% of the loss under low error variance ($\sigma_0 = 0.25$) and 74.1% of the loss under moderate error variance ($\sigma_0 = 0.5$). When mean error due to bias is large ($\beta = 0.25$), the bias-aware algorithm is able to recover 85.2% of the loss under low error variance and 55.5% of the loss under moderate error variance. The figure also suggests that an increase in σ_0 exacerbates the adverse impact of an increase

Figure 3. Percentage of Lost Life Years Recovered per Patient as a Function of Bias Factor for Various Levels of Error Variance Under a Bias-Aware Algorithm ($AUC_c = 0.656, AUC_i = 0.820$)



in β on the value of the bias-aware algorithm and vice versa.

The results of this subsection imply that (i) mitigating the impact of clinical-risk-information-induced bias through a bias-aware algorithm alone is likely to be challenging when the radiologist's bias-related behavior is highly unpredictable, and (ii) a large variation in biased interpretation of mammograms combined with a high mean error (or bias factor) is highly detrimental to the usefulness of a bias-aware algorithm in the breast cancer context. Therefore, when the bias-induced mean error and error variance are high, other bias-reduction efforts such as radiologist training and education become especially valuable.

5.5. Should the Clinical-Risk Information Be Used for Breast Cancer Diagnosis?

As shown in Proposition 3, not using the clinical-risk information for diagnosing breast cancer is sometimes better than using it if it will bias the radiologist. Table 2 shows the thresholds for the mean error or error variance due to bias above which using the clinical-risk information leads to an inferior expected utility or *AUC*. Therefore, in such cases, it is better to provide the clinical-risk information to neither the radiologist nor the referring physician. Mathematically, for a given β (or σ_0), we find the maximum σ_0 or (β) where $U^*(\beta, \sigma_0) \leq U(0,0,0,k(\alpha^*(0,0),0,0))$ holds. The table shows that, ceteris paribus, the threshold value for error mean decreases when the *AUC* of mammogram increases or when the *AUC* of the clinical-risk information decreases. Moreover, for a given pair of mammogram and clinical-risk information *AUC*s, the threshold for the error variance decreases with increasing error mean due to bias. The table emphasizes the critical role played by the relative discriminative ability of the clinical-risk information and mammogram in decisions concerning the use of clinical-risk information for cancer diagnosis. In particular, as the mammography technology improves relative to the clinical-risk information, even a small bias may make the use of clinical-risk information harmful for cancer screening.

Table 2. Thresholds for the Mean or Variance of Error for Nonusefulness of Clinical-Risk Information

(AUC_c, AUC_i)	$\beta = 0.1$ Allowable σ_0	$\beta = 0.25$ Allowable σ_0	$\sigma_0 = 1$ Allowable β
(0.656, 0.780)	1.82	0.72	0.20
(0.656, 0.820)	1.32	0.55	0.14
(0.656, 0.890)	0.79	0.32	0.08
(0.771, 0.820)	4.43	1.78	0.46
(0.771, 0.890)	2.01	0.81	0.20

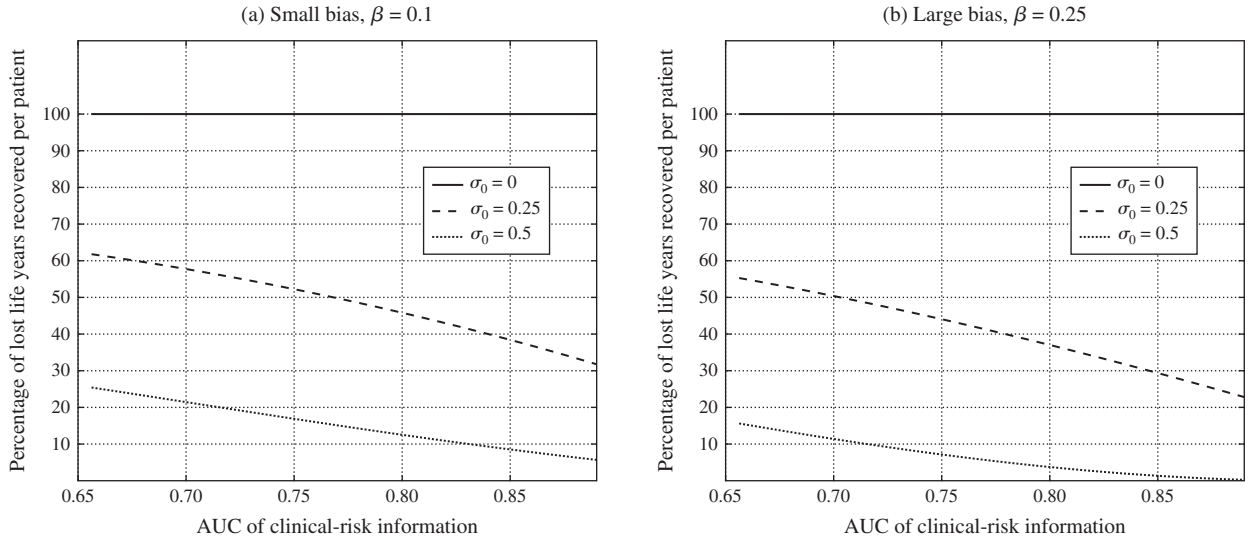
Notes. We choose $AUC_c = 0.771$ to represent the case that discriminative ability of clinical-risk information is improved. The *AUC* performance is the midpoint between the typical *AUC* for the clinical-risk information ($AUC_c = 0.656$) and the high mammography performance ($AUC_i = 0.890$).

5.6. Impact of Bias When the Accuracy of Clinical-Risk Information Improves

The well-established risk factors for breast cancer include factors related to personal characteristics and those related to family history. Genetic information, a factor related to family history, is increasingly used in risk-prediction models (Tyrer et al. 2004), as evidenced by the inclusion of them in diagnostic guidelines (Smith et al. 2013). Therefore, it is reasonable to expect that advances in genetic research will bring about substantial improvements in the discriminative ability of the clinical-risk information in the future. To examine its potential impact, we draw the percentage of lost life years recovered per patient by a bias-aware algorithm, as was defined in Section 5.4, versus the *AUC* of the clinical-risk information in Figure 4. We vary the *AUC* of clinical-risk information from 0.650 to 0.890 for a fixed mammogram *AUC* of 0.890.

Figure 4 offers several insights. As the *AUC* of clinical-risk information increases, the percentage of loss that can be recovered by accounting for bias decreases. Moreover, the recovery rate is smaller when the mean error or error variance because bias is higher. Therefore, correcting for the detrimental effects of bias is more difficult at higher clinical-risk information *AUC* values. In addition, our computational experiments suggest that absolute loss due to bias also diminishes (values not presented) when the *AUC* of clinical-risk information increases. In other words, when the information content of the clinical-risk information increases, the benefit from a bias-aware algorithm that

Figure 4. Percentage of Lost Life Years Recovered as a Function of Increasing AUC_c ($AUC_i = 0.820$)



accounts for bias diminishes. Hence, an important takeaway is that bias becomes less of an issue and a bias-aware algorithm becomes less valuable as the discriminative ability of the clinical-risk information improves relative to that of the mammography.

6. Discussion and Conclusion

When classification algorithms use human-generated input data that suffer from human biases, the predictions they generate may exacerbate the errors stemming from such biases. In the breast cancer context, a physician would provide the follow-up recommendation to a patient based on clinical-risk information and a radiologist's mammogram assessment, which could be biased by the patient's clinical-risk information. We examine the design and value of a bias-aware algorithm for use in a CDSS for the breast-cancer-diagnosis context. We show that the optimal bias-aware algorithm can eliminate the adverse impact of bias if there is no variability associated with the bias-induced error in the radiologist's assessment. On the other hand, in the presence of variability in the bias-induced error, the impacts of bias can be mitigated, but not eliminated, even if the algorithmic design is adjusted to account for the bias. The optimal bias-aware algorithm assigns a smaller (larger) weight to the clinical-risk information (radiologist's mammogram assessment) when the mean error increases, but the reverse happens when the error variance increases. Interestingly, under some conditions, a bias-aware algorithm may assign a larger weight to the clinical-risk information in the presence of bias than in its absence, even though the bias causes the mammogram assessment to already capture part of the clinical-risk information. Furthermore, we identify scenarios where not using the clinical-risk information is better than using it for breast cancer diagnosis, in case the clinical-risk information can potentially bias the radiologist. Using estimates obtained from a breast-cancer-outcomes database based on clinical-risk information and from the medical literature, we show that a bias-aware algorithm can significantly improve the expected patient life years. The magnitude of improvement depends critically on the relative discriminative abilities of clinical-risk information and mammogram.

6.1. Translating the Findings Into Clinical Practice

Realizing the gains indicated by our study depends on (i) finding ways to incorporate proper weighting of clinical-risk information and mammogram information and (ii) reducing, eliminating, or properly adjusting for the bias due to clinical-risk information. The former objective can be achieved by accounting for cognitive biases in the algorithmic design of decision aids such as CDSSs and facilitating their use by physicians. It is worth emphasizing that the radiologist's primary task is to assess the mammogram and communicate these findings to the referring physician for a final decision. The referring physician will likely know the patient's clinical history and patient preferences better. Therefore, it is reasonable that, while making the final decision for the patient, the physician should use an expert system such as a CDSS that accounts for the unintended over- or underassessment of risk because of the available information. Multiple studies have demonstrated how CDSSs informed by evidence-based medicine could enable better decisions and reduce inappropriate use of imaging and related services (Raja et al. 2012, Ip et al. 2013, Blackmore et al. 2011). Furthermore, with the error rates reaching almost 30% in imaging interpretations and cognitive biases being a significant reason, corrective actions through CDSSs can help reduce unnecessary testing or

improve detection (Lee et al. 2013). IS researchers have also recently called upon designers of decision support systems for "algorithmic accountability" (Ransbotham et al. 2016, p. 844) such that the design of algorithms explicitly accounts for biases of prior decision makers that exist in databases (Barocas and Selbst 2016). This paper provides a formal approach to account for a specific cognitive bias in the decision-making framework of the radiologist–physician pair and to design a CDSS algorithm in the context of breast cancer diagnosis. Our computational experiments show the feasibility and the value of such an algorithm.

There are many ways to achieve the latter objective. Although no solution will completely debias doctors, promoting evidence-based medicine, educating physicians about potential biases, and, most importantly, using decision-aid tools could help alleviate the detrimental effects of bias in the mammography process (Bornstein and Emler 2001). In addition, providing the clinical-risk information after the mammography interpretation could lessen if not totally eliminate patient mismanagement due to bias (Griscom 2002).

One challenge that a physician using a CDSS will likely face is that the radiologists she deals with could be biased to different degrees, both in terms of bias factor (i.e., mean error) and error variance. Consequently, a one-size-fits-all approach that implements the same aggregation and classification algorithm regardless of which radiologist provides the mammogram assessment will diminish the value of the CDSS. However, the relative simplicity of our model and solution approach makes it possible to incorporate the model into a decision-aid tool. Such a tool can be customized for the individual physicians, taking into account the individual biases of radiologists they work with. Such an effort would require prospective data collection and statistical analysis to determine the mean error or error variance due to the biased interpretations of individual radiologists.[11] Another issue to consider regarding the implementation of the bias-aware algorithm in a CDSS is the referring physician's attitude toward its predictions. Recent literature reports varying behavioral responses in a general setting. In particular, experiments show that people may lose confidence in algorithms that err—a phenomenon called algorithm aversion (Dietvorst et al. 2015). Moreover, empowering decision makers in overruling the algorithm and more flexibility in decision making can help mitigate the low trust in imperfect algorithms (Dietvorst et al. 2016). In the context of a bias-aware algorithm, the improved performance as a result of accounting for bias and the flexibility in decision making may engender increased trust. Yet, the inevitable imperfections in algorithmic predictions can erode trust, as with any other decision support system. How behavioral responses such as algorithm aversion would play out in the context of a bias-aware algorithm is uncertain and is a subject for future research.

The actual CDSS implementation would take significant effort and commitment, as was largely described in the health IT literature (Sim et al. 2001). Successful CDSS implementations in the past (Greenes 2014), especially those in the radiology domain (Li et al. 2013), are encouraging, suggesting the feasibility of translating our findings into a CDSS that can be used in practice. Such an effort would require further research and determination of parameters specific to a group of radiologists. A large health system that houses the radiology function and has advanced health IT capabilities could be a test bed for mammography quality improvement efforts around the algorithmic design (Britton 2015, e.g., for quality improvement efforts at a large system, see). Continued research into a contextual design is necessary to meet the CDSS's potential (O'Malley 2011). Particularly, future research could consider interactions among skewed decisions, design of CDSSs, and performance impact (Arnott 2006, Kohli and Piontek 2008). Our research is a step in that direction.

Note that our framework does not intend to debias the radiologist herself. Rather, it aims to debias the data generated by a biased radiologist and help the referring physicians who decide on the final course of action. In fact, prescriptively debiasing a biased radiologist is challenging and may be an infeasible task, as it will create more conflicts in radiologists' decision making (e.g., telling a radiologist to adjust their decision because of a biased assessment via a CDSS is unlikely to gain their buy-in). Also, the CDSS does not replace the radiologist; rather, it takes the radiologist's probabilistic assessment as input. We envision that the referring physician can have access to clinical-risk information, the radiologist's recommendation, other patient-related pertinent information (e.g., insurance information or other pressing concerns), and the CDSS's recommendation. The recommendations of the CDSS would be more useful in instances where the radiologist recommendation falls in the gray area such that the physician may or may not recommend a follow-up.[12] In mammography practice, this is often observed when the radiologist assesses a low to moderate risk of cancer that is typically managed with either a short-term imaging (within six month instead of regular screening) or further imaging or assessment. A finding of a cyst (not cancer) on a mammogram would be an example. The referring physician would then benefit from the CDSS's recommendation while evaluating the patient's overall health information, and could save the patient unnecessary procedures (or properly follow up the patient) with the help of the CDSS.

6.2. Future Directions

Our research can be extended in several directions. On the theoretical side, instead of restricting the analysis

to two attributes and two classification outcomes, the model can be extended to a general scenario with n attributes and m classification outcomes. The challenge with this extension lies in identifying and modeling all potential biases that exist among all n attributes. We also considered a linear-aggregation model. While such a model has been shown to be effective in different contexts, nonlinear aggregation models can provide additional insights into the impact of bias in other settings. The modeling of bias can also be examined further. For instance, in our model, the bias factor is the same whether the secondary attribute exceeds the mean or falls short of it. These coefficients can be different in a more general model. On the practice side, the lack of patient-level mammography data forced us to impute the parameters of the mammogram-related risk distribution using the clinical-risk distribution data. Patient-level mammography data would provide better estimates about the value of a bias-aware algorithm in the breast cancer context. In the same vein, little research quantifies the extent of bias in radiologists' assessments. Estimating the bias parameters requires extensive well-designed experiments. Finally, our model should be validated by implementing it in a CDSS in practice and comparing the actual results with the model predictions. Such an effort will help improve the model.

Acknowledgments

The authors thank Giri Kumar Tayi, senior editor; Sam Ransbotham, associate editor; and two anonymous reviewers for their constructive and helpful suggestions. The discussions during the authors' presentations at the departmental seminars of the University of Chicago and the University of Texas at Austin and conference presentations at 2016 INFORMS Annual Meeting (Nashville, TN) and 2017 POMS Annual Conference (Seattle, WA) were also beneficial.

Endnotes

[1] In this paper, we focus on the case where the radiologist has access to the same information as the referring physician; that is, they work for the same healthcare organization, and the information is available to all. According to the most recent figures, in-house radiology comprises about 60% of the radiology market Grand View Research (2016).

[2] We provide a discussion of different types of biases reflective of our setup in Section 2.1.

[3] Linear classifiers are used widely in many contexts including breast cancer diagnosis (Ayvaci et al. 2014, Chhatwal et al. 2009). In addition to being simple and allowing easier interpretation, the linear models have been found to perform at comparable levels relative to more sophisticated and complex nonlinear models (Ayer et al. 2010b).

[4] The CDSS is unlikely to consider every relevant factor available to the physician before making a recommendation. We provide further considerations on realizing the gains from the CDSS and translating it into practice in Section 6.1.

[5] Mammography interpretation is highly standardized. In practice, radiologists report their assessments of cancer probability using seven main BI-RADS assessment categories and three subcategories (Eberl et al. 2006). Each of the categories and subcategories corresponds to a probability of cancer associated with the mammogram.

[6] In Section 4.5, we examine a special, and possibly also realistic, case of this model in which the mean error and error variance are interdependent in the sense that an increase in bias factor, β, increases both absolute mean error and error variance. With only a single source for error (the radiologist), as in the mammogram context, the mean error and error variance could be interdependent.

[7] Theoretically, α can take any real value. In practice, when each piece of risk information is informative in the presence of the other, the optimal α will lie in the interval $[0, 1]$ in the absence of bias.

[8] Later in the numerical study, we find that the standard deviations belonging to cancer and healthy patients for clinical-risk information are similar, consistent with the assumption.

[9] Note that the high intra- and interradiologist variability reported in mammography interpretation suggests that high bias-induced errors are a common phenomenon (Elmore et al. 1994), making the described case a possible scenario.

[10] Approximately 39 million women are screened in a year (U.S. Food and Drug Administration 2015).

[11] For example, Adomavicius et al. (2013) quantified anchoring bias in the context of recommender systems for user preference ratings for TV shows.

[12] When the radiologist clearly recommends a follow-up based on highly suspicious findings, the CDSS will likely concur because of the expected higher weight on the mammogram information.

References

Adomavicius G, Zhang J (2016) Classification, ranking, and top-K stability of recommendation algorithms. *INFORMS J. Comput.* 28(1):129–147.

Adomavicius G, Bockstedt J, Curley S, Zhang J (2014) De-biasing user preference ratings in recommender systems. Tintarev N, O'Donovan J, Brusilovsky P, Felfernig A, Semeraro G, Lops P, eds. *Joint Workshop Interfaces Human Decision Making Recommender Systems, October 6, Foster City, CA*, 2–9.

Adomavicius G, Bockstedt JC, Curley SP, Zhang J (2013) Do recommender systems manipulate consumer preferences? A study of anchoring effects. *Inform. Systems Res.* 24(4):956–975.

Agarwal R, Dhar V (2014) Big data, data science, and analytics: The opportunity and challenge for IS research. *Inform. Systems Res.* 25(3):443–448.

Alpert HR, Hillman BJ (2004) Quality and variability in diagnostic radiology. *J. Amer. College Radiology* 1(2):127–132.

American Cancer Society (2014) Cancer facts and figures 2014. Technical report, American Cancer Society, Atlanta.

American College of Radiology (1998) *Illustrated Breast Imaging Reporting and Data System (BI-RADS TM)*, 3rd ed. (American College of Radiology, Reston, VA).

Arnott D (2006) Cognitive biases and decision support systems development: A design science approach. *Inform. Systems J.* 16(1): 55–78.

Ayer T, Ayvaci MUS, Liu ZX, Alagoz O, Burnside ES (2010a) Computer-aided diagnostic models in breast cancer screening. *Imaging Medicine* 2(3):313–323.

Ayer T, Chhatwal J, Alagoz O, Kahn CE Jr, Woods RW, Burnside ES (2010b) Comparison of logistic regression and artificial neural network models in breast cancer risk estimation. *Radiographics* 30(1):13–22.

Ayvaci MUS, Ahsen ME, Raghunathan S, Gharibi Z (2017) Timing the use of breast cancer risk information in biopsy decision-making. *Production Oper. Management* 26(7):1333–1358.

Ayvaci MUS, Alagoz O, Ahsen ME, Burnside E (2018) Preference-sensitive management of post-mammography decisions in breast cancer diagnosis. *Production Oper. Management* 27(12):2313–2338.

Ayvaci MUS, Alagoz O, Chhatwal J, del Rio AM, Sickles EA, Nassif H, Kerlikowske K, et al. (2014) Predicting invasive breast cancer versus DCIS in different age groups. *BMC Cancer* 14(1):584.

Bai X, Padman R, Ramsey J, Spirtes P (2008) Tabu search-enhanced graphical models for classification in high dimensions. *INFORMS J. Comput.* 20(3):423–437.

Ballard-Barbash R, Taplin SH, Yankaskas BC, Ernster VL, Rosenberg RD, Carney PA, Barlow WE, et al. (1997) Breast cancer surveillance consortium: A national mammography screening and outcomes database. *Amer. J. Roentgenology* 169(4):1001–1008.

Bansal S, Gutierrez GJ, Keiser JR (2017) Using experts' quantile judgments to quantify uncertainties for new product development. *Oper. Res.* 65(5):1115–1130.

Barlow WE, White E, Ballard-Barbash R, Vacek PM, Titus-Ernstoff L, Carney PA, Tice JA, et al. (2006) Prospective breast cancer risk prediction model for women undergoing screening mammography. *J. National Cancer Inst.* 98(17):1204–1214.

Barocas S, Selbst AD (2016) Big data's disparate impact. *California Law Rev.* 104:671–732.

Berg WA, Campassi C, Langenberg P, Sexton MJ (2000) Breast imaging reporting and data system: Inter- and intraobserver variability in feature analysis and final assessment. *Amer. J. Roentgenology* 174(6):1769–1777.

Blackmore CC, Mecklenburg RS, Kaplan GS (2011) Effectiveness of clinical decision support in controlling inappropriate imaging. *J. Amer. College Radiology* 8(1):19–25.

Blumenthal-Barby JS, Krieger H (2015) Cognitive biases and heuristics in medical decision making: A critical review using a systematic search strategy. *Medical Decision Making* 35(4):539–557.

Bornstein BH, Emler AC (2001) Rationality in medical decision making: A review of the literature on doctors' decision-making biases. *J. Evaluation Clinical Practice* 7(2):97–107.

Boylu F, Aytug H, Koehler GJ (2007) Using a genetic algorithm to solve the strategic learning problem. Technical report, Decision and Information Sciences, University of Florida, Gainesville.

Boylu F, Aytug H, Koehler GJ (2010) Induction over strategic agents. *Inform. Systems Res.* 21(1):170–189.

Brennan T, Dieterich W, Ehret B (2009) Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice Behav.* 36(1):21–40.

Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, Samsa G, et al. (2012) Effect of clinical decision-support systems: A systematic review. *Ann. Internal Medicine* 157(1):29–43.

Britton M (2015) Improving care through mammogram quality research. Accessed June 13, 2017, https://intermountainhealthcare.org/blogs/2015/10/breast-cancer-awareness-month-research-latest.

Burnside ES, Davis J, Chhatwal J, Alagoz O, Lindstrom MJ, Geller BM, Littenberg B, et al. (2009) Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings. *Radiology* 251(3):663–672.

Campbell SD, Sharpe SA (2009) Anchoring bias in consensus forecasts and its effect on market prices. *J. Financial Quant. Anal.* 44(2):369–390.

Cantor SB, Sun CC, Tortolero-Luna G, Richards-Kortum R, Follen M (1999) A comparison of C/B ratios from studies using receiver operating characteristic curve analysis. *J. Clinical Epidemiology* 52(9):885–892.

Carney PA, Cook AJ, Miglioretti DL, Feig SA, Bowles EA, Geller BM, Kerlikowske K, et al. (2012) Use of clinical history affects accuracy of interpretive performance of screening mammography. *J. Clinical Epidemiology* 65(2):219–230.

Cavusoglu H, Mishra B, Raghunathan S (2005) The value of intrusion detection systems in information technology security architecture. *Inform. Systems Res.* 16(1):28–46.

Chen JH, Asch SM (2017) Machine learning and prediction in medicine—Beyond the peak of inflated expectations. *New England J. Medicine* 376(26):2507–2509.

Chhatwal J, Alagoz O, Lindstrom MJ, Kahn CE, Shaffer KA, Burnside ES (2009) Original research. A logistic regression model based on the national mammography database format to aid breast cancer diagnosis. *Amer. J. Roentgenology* 192(4): 1117–1127.

Clark P, Niblett T (1989) The CN2 induction algorithm. *Machine Learn.* 3(4):261–283.

Clemen RT, Winkler RL (1999) Combining probability distributions from experts in risk analysis. *Risk Anal.* 19(2):187–203.

Coiera E (2015) Technology, cognition and error. *BMJ Quality Safety* 24(7):417–422.

Croskerry P (2003) The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine* 78(8): 775–780.

Dalvi N, Domingos P, Sanghai S, Verma D (2004) Adversarial classification. *Proc. Tenth ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 99–108.

Dekel O, Shamir O, Xiao L (2010) Learning to classify with missing and corrupted features. *Machine Learn.* 81(2):149–178.

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Experiment. Psych.: General* 144(1):114–126.

Dietvorst BJ, Simmons JP, Massey C (2016) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Sci.* 64(3):1155–1170

Eberl MM, Fox CH, Edge SB, Carter CA, Mahoney MC (2006) BI-RADS classification for management of abnormal mammograms. *J. Amer. Board Family Medicine* 19(2):161–164.

Elliott C (2017) New artificial intelligence promises to make travel a little smarter. Does it? *Washington Post* (June 22), http://wapo.st/2t0wCjX?tid=ss_tw&utm_term=.acde7d2bff94.

Elmore JG, Wells CK, Howard DH, Feinstein AR (1997) The impact of clinical history on mammographic interpretations. *J. Amer. Medical Assoc.* 277(1):49–52.

Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR (1994) Variability in radiologists' interpretations of mammograms. *New England J. Medicine* 331(22):1493–1499.

Englich B, Soder K (2009) Moody experts—How mood and expertise influence judgmental anchoring. *Judgment Decision Making* 4(1):41–50.

Engmann NJ, Golmakani MK, Miglioretti DL, Sprague BL, Kerlikowske K (2017) Population-attributable risk proportion of clinical risk factors for breast cancer. *J. Amer. Medical Assoc. Oncology* 3(9):1228–1236.

Eraker SA, Politser P (1982) How decisions are reached: Physician and patient. *Ann. Internal Medicine* 97(2):262–268.

Faraggi D, Reiser B (2002) Estimation of the area under the ROC curve. *Statist. Medicine* 21(20):3093–3106.

Fisher M, Raman A (1996) Reducing the cost of demand uncertainty through accurate response to early sales. *Oper. Res.* 44(1):87–99.

Freedman AN, Seminara D, Gail MH, Hartge P, Colditz GA, Ballard-Barbash R, Pfeiffer RM (2005) Cancer risk prediction models: A workshop on development, evaluation, and application. *J. National Cancer Inst.* 97(10):715–723.

Fryback DG, Stout NK, Rosenberg MA, Trentham-Dietz A, Kuruchittham V, Remington PL (2006) The Wisconsin breast cancer epidemiology simulation model. *J. National Cancer Inst. Monographs* 2006(36):37–47.

Gartner D, Kolisch R, Neill DB, Padman R (2015) Machine learning approaches for early DRG classification and resource allocation. *INFORMS J. Comput.* 27(4):718–734.

Goes PB (2013) Editor's comments: Information systems research and behavioral economics. *Management Inform. Systems Quart.* 37(3):iii–viii.

Gold MR (1996) *Cost-Effectiveness in Health and Medicine* (Oxford University Press, New York).

Gram IT, Lund E, Slenker SE (1990) Quality of life following a false positive mammogram. *British J. Cancer* 62(6):1018–1022.

Grand View Research (2016) Teleradiology market size & forecast report, 2013–2024. Technical report, Grand View Research, San Francisco.

Greenes R (2014) Definition, scope, and challenges. Greenes RA, ed. *Clinical Decision Support: The Road Ahead* (Academic Press, Cambridge, MA), 3–47.

Griscom NT (2002) A suggestion: Look at the images first, before you read the history. *Radiology* 223(1):9–10.

Grushka-Cockayne Y, Jose VRR, Lichtendahl KC Jr (2016) Ensembles of overfit and overconfident forecasts. *Management Sci.* 63(4):1110–1130.

Gunderman RB (2009) Biases in radiologic reasoning. *Amer. J. Roentgenology* 192(3):561–564.

Hajian S, Bonchi F, Castillo C (2016) Algorithmic bias: From discrimination discovery to fairness-aware data mining. Krishnapuram B, ed. *Proc. 22nd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 2125–2126.

Hawley JR, Taylor CR, Cubbison AM, Erdal BS, Yildiz VO, Carkaci S (2016) Influences of radiology trainees on screening mammography interpretation. *J. Amer. College Radiology* 13(5):554–561.

Haybittle JL (1998) Life expectancy as a measurement of the benefit shown by clinical trials of treatment for early breast cancer. *Clinical Oncology* 10(2):92–94.

Holmes-Rovner M, Nelson WL, Pignone M, Elwyn G, Rovner DR, O'Connor AM, Coulter A, et al. (2007) Are patient decision aids the best way to improve clinical decision making? Report of the IPDAS symposium. *Medical Decision Making* 27(5):599–608.

Hong T, Tsang S (1997) A generalized version space learning algorithm for noisy and uncertain data. *IEEE Trans. Knowledge Data Engrg.* 9(2):336–340.

Ichikawa LE, Barlow WE, Anderson ML, Taplin SH, Geller BM, Brenner RJ (2010) Time trends in radiologists' interpretive performance at screening mammography from the community-based breast cancer surveillance consortium, 1996–2004. *Radiology* 256(1):74–82.

Ip IK, Schneider L, Seltzer S, Smith A, Dudley J, Menard A, Khorasani R (2013) Impact of provider-led, technology-enabled radiology management program on imaging. *Amer. J. Medicine* 126(8):687–692.

Jiang Z, Mookerjee VS, Sarkar S (2005) Lying on the web: Implications for expert systems redesign. *Inform. Systems Res.* 16(2):131–148.

Kao H, Tang K (2013) Cost-sensitive decision tree induction with label-dependent late constraints. *INFORMS J. Comput.* 26(2): 238–252.

Kardaras C, Žitković G (2011) Stability of the utility maximization problem with random endowment in incomplete markets. *Math. Finance* 21(2):313–333.

Kohli R, Piontek F (2008) DSS in healthcare: Advances and opportunities. Burstein F, Holsapple CW, eds. *Handbook on Decision Support Systems*, Vol. 2 (Springer, Berlin), 483–497.

Lee CS, Nagy PG, Weaver SJ, Newman-Toker DE (2013) Cognitive and system factors contributing to diagnostic errors in radiology. *Amer. J. Roentgenology* 201(3):611–617.

Li KC, Marcovici P, Phelps A, Potter C, Tillack A, Tomich J, Tridandapani S (2013) Digitization of medicine: How radiology can take advantage of the digital revolution. *Academic Radiology* 20(12):1479–1494.

Li M, Petruzzi NC, Zhang J (2017) Overconfident competing newsvendors. *Management Sci.* 63(8):2637–2646.

Lin A, Harris M, Zalis M (2010) Initial observations of electronic medical record usage during CT and MRI interpretation: Frequency of use and impact on workflow. *Amer. J. Roentgenology* 195(1):188–193.

Lo JY, Baker JA, Kornguth PJ, Floyd CE (1999) Effect of patient history data on the prediction of breast cancer from mammographic findings with artificial neural networks. *Academic Radiology* 6(1):10–15.

Loy CT, Irwig L (2004) Accuracy of diagnostic tests read with and without clinical information: A systematic review. *J. Amer. Medical Assoc.* 292(13):1602–1609.

MacCormack A, Verganti R (2003) Managing the sources of uncertainty: Matching process and context in software development. *J. Product Innovation Management* 20(3):217–232.

Mamede S, van Gog T, van den Berge K, Rikers RMJP, van Saase JLCM, van Guldener C, Schmidt HG (2010) Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. *J. Amer. Medical Assoc.* 304(11):1198–1203.

Mandelblatt JS, Wheat ME, Monane M, Moshief RD, Hollenberg JP, Tang J (1992) Breast cancer screening for elderly women with and without comorbid conditions. *Ann. Internal Medicine* 116(9):722–730.

Metz CE (1978) Basic principles of ROC analysis. *Seminars Nuclear Medicine* 8(4):283–298.

Meyer G, Adomavicius G, Johnson PE, Elidrisi M, Rush WA, Sperl-Hillen JM, O'Connor PJ (2014) A machine learning approach to improving dynamic decision making. *Inform. Systems Res.* 25(2):239–263.

Mookerjee VS (2001) Debiasing training data for inductive expert system construction. *IEEE Trans. Knowledge Data Engrg.* 13(3):497–512.

Nance JW Jr, Meenan C, Nagy PG (2013) The future of the radiology information system. *Amer. J. Roentgenology* 200(5):1064–1070.

National Cancer Institute (2014a) Breast cancer risk assessment tool. Accessed January 15, 2015, http://www.cancer.gov/bcrisktool/.

National Cancer Institute (2014b) Breast Cancer PDQ® screening. Accessed January 15, 2015, http://www.cancer.gov/cancertopics/pdq/screening/breast/healthprofessional/page7.

Obermeyer Z, Lee TH (2017) Lost in thought–the limits of the human mind and the future of medicine. *New England J. Medicine* 377(13):1209–1211.

Ogdie AR, Reilly JB, Pang WG, Keddem S, Barg FK, Von Feldt JM, Myers JS (2012) Seen through their eyes: Residents' reflections on the cognitive and contextual components of diagnostic errors in medicine. *Academic Medicine* 87(10):1361–1367.

O'Malley AS (2011) Tapping the unmet potential of health information technology. *New England J. Medicine* 364(12):1090–1091.

Pace LE, Keating NL (2014) A systematic assessment of benefits and risks to guide breast cancer screening decisions. *J. Amer. Medical Assoc.* 311(13):1327–1335.

Parssian A (2006) Managerial decision support with knowledge of accuracy and completeness of the relational aggregate functions. *Decision Support Systems* 42(3):1494–1502.

Pasquale F (2015) Digital star chamber. *Aeon* (August 18), https://aeon.co/essays/judge-jury-and-executioner-the-unaccountable-algorithm.

Patel VL, Kannampallil TG, Shortliffe EH (2015) Role of cognition in generating and mitigating clinical errors. *BMJ Quality Safety* 24(7):468–474.

Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, Conant EF, et al. (2005) Diagnostic performance of digital versus film mammography for breast-cancer screening. *New England J. Medicine* 353(17):1773–1783.

Provost F (2005) Toward economic machine learning and utility-based data mining. Weiss G, Saar-Tsechansky M, Zadrozny B, eds. *Proc. 1st Internat. Workshop Utility-Based Data Mining* (Association for Computing Machinery, New York), 1.

Quinlan JR (1986) The effect of noise on concept learning. Michalski RS, Carbonell JG, Mitchell TM, eds. *Machine Learning: An Artificial Intelligence Approach*, Vol. 2, 149–166.

Raja AS, Ip IK, Prevedello LM, Sodickson AD, Farkas C, Zane RD, Hanson R, et al. (2012) Effect of computerized clinical decision support on the use and yield of CT pulmonary angiography in the emergency department. *Radiology* 262(2):468–474.

Ranjan R, Gneiting T (2010) Combining probability forecasts. *J. Roy. Statist. Soc.: Ser. B* (*Statist. Methodology*) 72(1):71–91.

Ransbotham S, Fichman RG, Gopal R, Gupta A (2016) Special section introduction—Ubiquitous IT and digital vulnerabilities. *Inform. Systems Res.* 27(4):834–847.

Rosenberg RD, Yankaskas BC, Abraham LA, Sickles EA, Lehman CD, Geller BM, Carney PA, et al. (2006) Performance benchmarks for screening mammography. *Radiology* 241(1):55–66.

Saar-Tsechansky M, Provost F (2007) Handling missing values when applying classification models. *J. Machine Learn. Res.* 8(7): 1624–1657.

Saposnik G, Redelmeier D, Ruff CC, Tobler PN (2016) Cognitive biases associated with medical decisions: A systematic review. *BMC Medical Informatics Decision Making* 16(1):138.

Sheng VS, Provost F, Ipeirotis PG (2008) Get another label? Improving data quality and data mining using multiple, noisy labelers. Liu B, Sarawagi S, Li Y, eds. *Proc. 14th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 614–622.

Sibbald M, Cavalcanti RB (2011) The biasing effect of clinical history on physical examination diagnostic accuracy. *Medical Ed.* 45(8):827–834.

Siegel R, Ma J, Zou Z, Jemal A (2014) Cancer statistics, 2014. *CA: Cancer J. Clinicians* 64(1):9–29.

Sim I, Gorman P, Greenes RA, Haynes RB, Kaplan B, Lehmann H, Tang PC (2001) Clinical decision support systems for the practice of evidence-based medicine. *J. Amer. Medical Informatics Assoc.* 8(6):527–534.

Smith RA, Brooks D, Cokkinides V, Saslow D, Brawley OW (2013) Cancer screening in the United States, 2013. *CA: Cancer J. Clinicians* 63(2):87–105.

Stone M (1961) The opinion pool. *Ann. Math. Statist.* 32(4): 1339–1342.

Tice JA, Cummings SR, Smith-Bindman R, Ichikawa L, Barlow WE, Kerlikowske K (2008) Using clinical factors and mammographic breast density to estimate breast cancer risk: Development and validation of a new predictive model. *Ann. Internal Medicine* 148(5):337–347.

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185(4157):1124–1131.

Tyrer J, Duffy SW, Cuzick J (2004) A breast cancer prediction model incorporating familial and personal risk factors. *Statist. Medicine* 23(7):1111–1130.

Ulvila JW, Gaffney JE Jr (2004) A decision analysis method for evaluating computer intrusion detection systems. *Decision Anal.* 1(1):35–50.

U.S. Food and Drug Administration (2015) MQSA national statistics. Accessed January 15, 2015, http://www.fda.gov/Radiation-EmittingProducts/MammographyQualityStandardsActandProgram/FacilityScorecard/ucm113858.htm.

Wallsten TS (1981) Physician and medical student bias in evaluating diagnostic information. *Medical Decision Making* 1(2):145–164.

Winkler RL (1981) Combining probability distributions from dependent information sources. *Management Sci.* 27(4):479–488.

Wu W, Zhang W, Li H (2003) Knowledge acquisition in incomplete fuzzy information systems via the rough set approach. *Expert Systems* 20(5):280–286.

Wu Y, Vanness DJ, Ayvaci MUS, Alagoz O, Burnside ES (2011) Expected utility model used to compare the value of screening versus diagnostic mammography. 33*rd Annual Meeting Soc. Medical Decision Making* (Society for Medical Decision Making, Hillsborough, NJ), 245–248.

Zhang J, Aytug H, Koehler GJ (2014) Discriminant analysis with strategically manipulated data. *Inform. Systems Res.* 25(3): 654–662.