Introduction to Biostatistical Machine Learning

Spencer Wozniak*

Department of Epidemiology and Biostatistics
Michigan State University
East Lansing, MI 48824, USA

*Corresponding author
Spencer Wozniak
(woznia79@msu.edu)

Motivation

I have a background in machine learning (ML) research, I am planning on becoming a physician, and I am especially interested in the integration of ML into the practice of medicine.

I saw there were many seminars on ML, so I wanted to use this Honors project as an opportunity to take a deep dive into the concepts that will likely be of relevance if/when I get into medical ML research in the future.

The overall goal of this project is to introduce the application of ML techniques in medical diagnostics and epidemiology. It attempts to introduce concepts and terms in the context of examples to facilitate a better understanding of the content. It is particularly suited to readers with a medical background.

Notes

All dichotomous variables mentioned in the paper are coded {0,1} for negative and positive results, respectively (e.g. a patient without CHF is coded 0 and a patient with CHF is coded 1).

I tried to stay away from technical mathematical details in this project, especially with advanced ML concepts, to focus on the overall significance of the methods and research. Details on fitting / training models are brief, if mentioned at all.

If you believe that I missed any important concepts, if you have any advice, or if there is any further reading that you think may be of interest to me, any comments/suggestions are much appreciated!

Lecture not included in paper

Note that I did watch 5 lectures, but I only wrote about 4 because I felt that the concepts of the final lecture did not add much to the overall paper. Also, I included a background section, which probably took longer than the rest of the paper combined and takes up as much space as writing about another lecture would.

The lecture not included is:

> Ganguli, Arkaprabha. (2023, October 12). *Deep Learning-Aided Feature Selection for Cognitive Reserve with Highly Correlated Diffusion-MRI Tractography Data.* [Lecture]. Argonne National Laboratory, Lemont, IL, USA. https://youtu.be/Ht6y7XErTCg

Abstract

Background

- Introduction to relevant concepts in medical diagnostics.
- Introduction to the general idea of ML in medical diagnostics.

Traditional Biostatistical ML Methods

- References study by De los Campos (2021) to explain how traditional biostatistical ML methods are applied, specifically in the realm of genetics.

Modern ML in Epidemiology

- References study by Petersen (2022) to explain the concepts of supervised and unsupervised in the context of ML, and to provide examples of their applications alongside traditional methods in epidemiology.

Modern ML in Medicine

- References study by Levin (2023) to provide an example of the application of a neural network in medical diagnostics, specifically the interpretation of histopathological specimens to classify prostate cancer.

Causal Estimation with ML

- References study by Tec (2024) to provide an example of the application of a neural network in epidemiology, along with a framework for establishing causation that doubles as a responsible approach to research and innovation.

Concluding Remarks

- Limitations of biostatistical ML
- Future of biostatistical ML

Keywords

Biostatistics, epidemiology, medicine, medical diagnosis, machine learning, artificial intelligence, classification, dichotomous variable, neural network, logistic regression, linear regression, odds ratio, relative risk, convolutional neural network

Background

Principles of Medical Diagnostics

In medicine, **sensitivity** (SN) and **specificity** (SP) are the fundamental metrics that quantify the accuracy of a diagnostic test. SN, the *true positive rate*, refers to the proportion of individuals with a specific outcome who are correctly identified by a test as having that outcome, whereas SP, the *true negative rate*, refers to the proportion of individuals without the outcome who are correctly identified by a test as not having that outcome. Diagnostic tests with high SN and SP are considered reliable because they minimize the number of false positives and false negatives [1].

For example, B-type natriuretic peptide (BNP) is a routine laboratory test used to screen for congestive heart failure (CHF) in medicine, as BNP is released into the bloodstream in response to ventricular volume expansion and pressure overload, two conditions commonly found in CHF [2]. In patients presenting to the emergency room with shortness of breath (SOB), using a cutoff BNP level of 150 pg/mL can distinguish CHF from other causes of SOB with a SN of 85% and SP of 83%, considerable accuracy for a relatively simple test [1].

Receiver Operating Characteristic (**ROC**) **curves** plot power, SN, against the false discovery rate (FDR), $1 - \text{SP}$, at various cutoff levels. These plots are instrumental in the evaluation of diagnostic tests, not only because they allow for identifying an optimal cutoff by depicting the tradeoff between power and FDR, but also because calculating the **Area Under the ROC Curve** (**AUC**) provides a measure of the test's overall diagnostic accuracy: an AUC close to 1.0 indicates excellent diagnostic performance, whereas an AUC closer to 0.5 suggests that diagnostic utility is no better than random chance. For reference, the AUC for BNP in classifying SOB as CHF-related is 0.91 [1], and the ROC curve is shown in Figure 1.

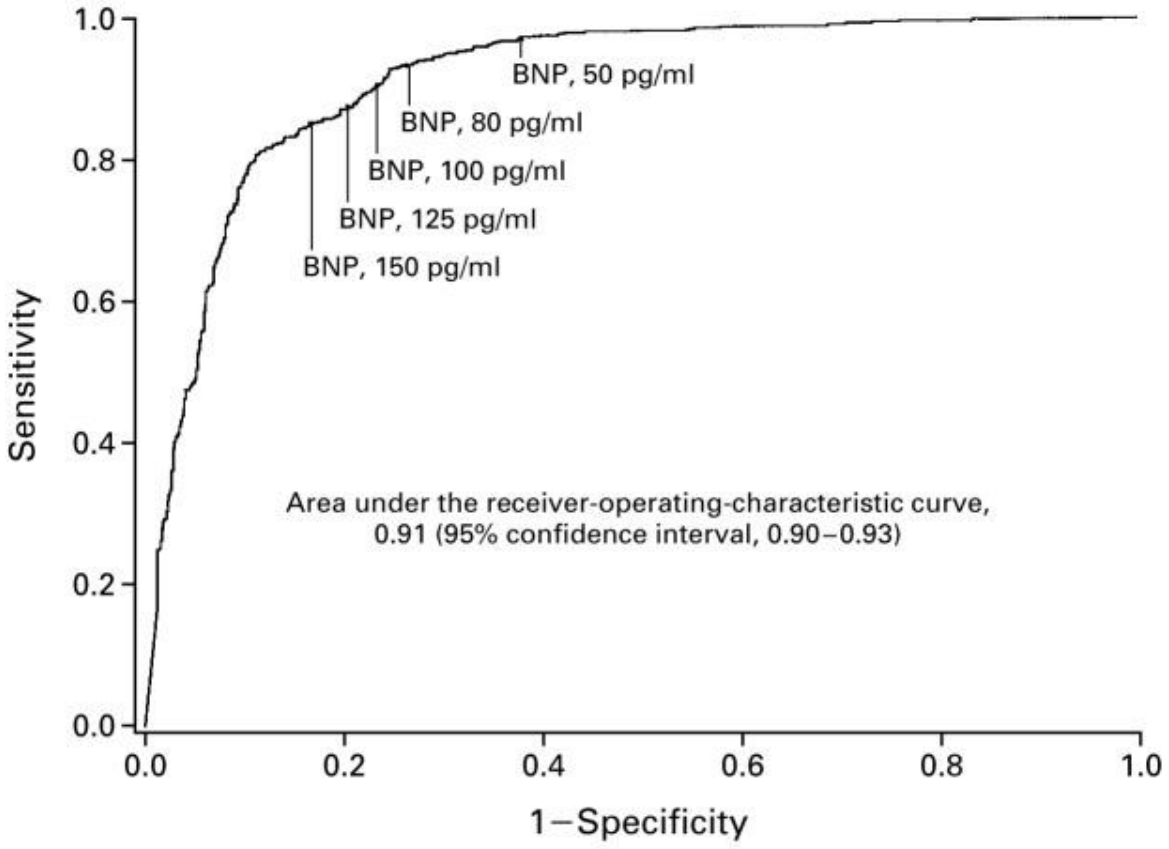

Figure 1. ROC curve for BNP in differentiating SOB due to CHF vs other causes. Adapted from [2].

Medical diagnostics can be summarized as mapping predictors to (probabilities of) outcomes; in this case, BNP is a single predictor of the outcome, CHF. However, the clinically accurate diagnosis of CHF (like most outcomes in the real world) relies on a broad set of predictors including overall clinical picture, electrocardiogram findings, and echocardiogram results [3]. Accordingly, this kind of analysis can be upscaled by designing machine learning (ML) models that incorporate many predictors. This is achieved by calculating SN and SP values for an overall model, then creating the ROC curve.

Odds represent the ratio of the probability that an event will occur to the probability that it will not. This is defined mathematically as:

$$\text{Odds} = \frac{p}{1-p}$$

where p is the probability that the event will occur. Based on this formula, if the probability of an event is 50%, the odds are 1, often expressed as "1-to-1" or "even odds," meaning there is an equal likelihood of the event occurring as not occurring.

In statistics, **logistic regression** is often utilized to estimate the probability of an outcome based on multiple predictors. It is one of the simplest ML methods, based on the logit function:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

which transforms a probability, p, into "log-odds." For instance, a logistic regression model that incorporates BNP, along with other relevant predictors, $x_2, \dots, x_n$, could be implemented for diagnosing CHF. Such a model would attempt to approximate the logit function:

$$\text{logit}\big(P(\text{CHF} = 1|\text{BNP}, x_2, \dots, x_n)\big) = \beta_0 + \beta_1 \times \text{BNP} + \beta_2 \times x_2 + \cdots + \beta_n \times x_n$$

where $\text{CHF} = 1$ indicates a positive CHF diagnosis. The model would "learn" by optimizing the parameters $\beta_0, \beta_1, \dots, \beta_n$ based on known predictors and outcomes within the **training data**.

In ML, rather than merely adjusting a lab value threshold, you could vary the probability threshold for classifying a dichotomous variable (in this case, the presence or absence of CHF) to generate an ROC curve. The probability of CHF given the predictors would be calculated via:

$$P(\text{CHF} = 1|\text{BNP}, x_2, \dots, x_n) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \times \text{BNP} + \beta_2 \times x_2 + \cdots + \beta_n \times x_n)}}$$

which effectively converts log-odds of having CHF, which is given by the optimized model based on a specific input set of predictors (i.e. a specific patient), into a probability. Having a probability

not only simplifies our interpretation of the model's outputs, but it allows for the calculation of SN and SP, thus allowing ROC analysis. A visual illustration of this process is shown in Figure 2. Note that in the real world, multiple thresholds may be considered simultaneously (e.g. negative, indeterminate, and positive results), but ROC analysis can be extended far beyond these cases.

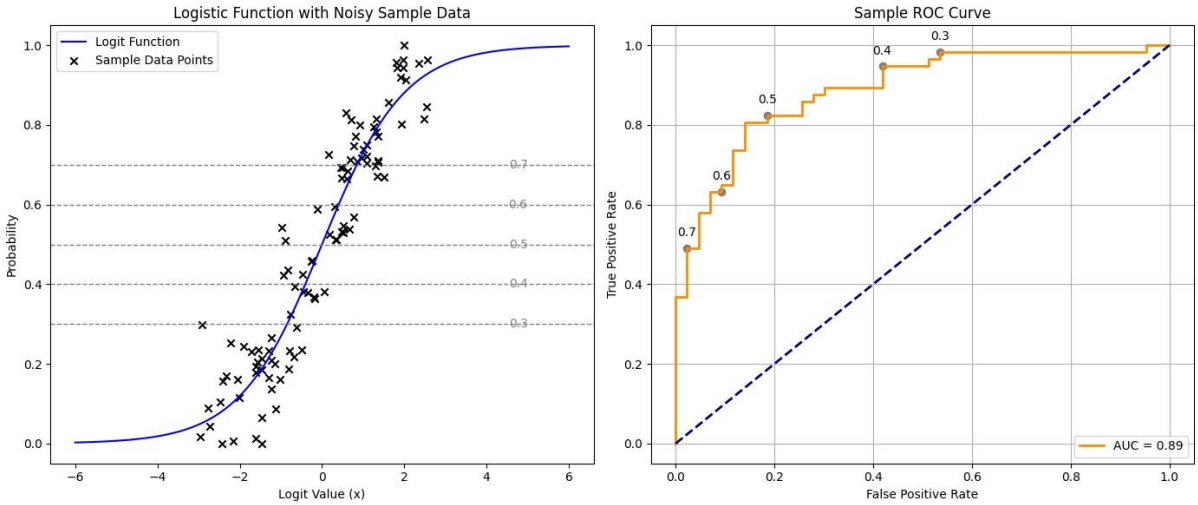

Figure 2. Logit function and ROC curve illustrating the idea of varying probability thresholds. On the left, the logit function is depicted with horizontal lines indicating different probability thresholds, illustrating how sample data points with relatively indeterminate model predictions ($p \in [0.3,0.7]$) could be classified as positive (above the threshold line) or negative (below the threshold line) depending on the value of the threshold. On the right, the sample ROC curve is shown with points representing the power and FDR at these varying probability thresholds. Sample data was simulated via normal distribution sampling and the plots were composed using the matplotlib package in Python.

A similar strategy could also be employed for more complex ML models, like **convolutional neural networks** (CNNs). For instance, a CNN could be designed to assess for CHF based on visual echocardiogram data [5], according to the simple equation:

$$\mathrm{CNN}(\mathrm{echo}) = P(\mathrm{CHF} = 1 \mid \mathrm{echo})$$

indicating the CNN processing an echocardiogram to predict the probability of having CHF, given that echocardiogram. With this kind of model, you could generate an ROC curve in the same fashion as with the logistic regression example: by varying the probability threshold that indicates the presence or absence of CHF.

Traditional Biostatistical ML Methods

Biostatisticians and epidemiologists often apply statistical models to identify relationships between predictors and outcomes. For example, Gustavo de los Campos, PhD, a professor in the Department of Epidemiology and Biostatistics at Michigan State University (2021) investigated the application of various statistical methods, outlined in Table 1, in genome-wide association studies (GWAS). GWAS aim to identify and localize genetic variants, especially single nucleotide polymorphisms (SNPs), that influence phenotypes like disease risk [6]. In this case, genetic sequence is the predictor (really many predictors – 1 for each base pair in the DNA sequence of interest), and disease risk is the outcome.

Table 1. Statistical methods described in [6].

Method	Description
SMR (Single Marker Regression)	Tests the statistical association between each predictor and outcome independently, one predictor at a time. This may involve linear or logistic regression. This method is effective for initial screening of predictors, but it is a relatively naïve approach overall.
LASSO (Least Absolute Shrinkage and Selection Operator)	Introduces a penalty term to shrink statistical coefficients (e.g. β_i) of less important predictors (e.g. x_i) to help reduce model complexity. This method is useful for eliminating predictors that do not contribute significantly to the model's predictive power, especially when the number of predictors far exceeds the number of predicted outcomes, like in GWAS.
BVS (Bayesian Variable Selection)	Incorporates probability distributions that reflect prior knowledge and uses Bayesian statistics to update these distributions based on training data. This method allows researchers to guide the model's focus towards associations that are more likely based on pre-existing knowledge.
Forward regression	A stepwise approach where predictors are sequentially added to a regression model (e.g. a logistic regression) in a way that results in the greatest increase in the fit of the model at each iteration. This method is useful when predictors interact because it considers each predictor's contribution in the context of the others already chosen.

To evaluate the efficacy of the different statistical methods, De los Campos used simulations in which the training data, composed of DNA sequences to predict a dichotomous outcome, was generated in a way that represents a condition genetically determined by 2 specific SNPs (known in this context as "causal SNPs") [6]. Selected results of this study are shown in Figure 2.

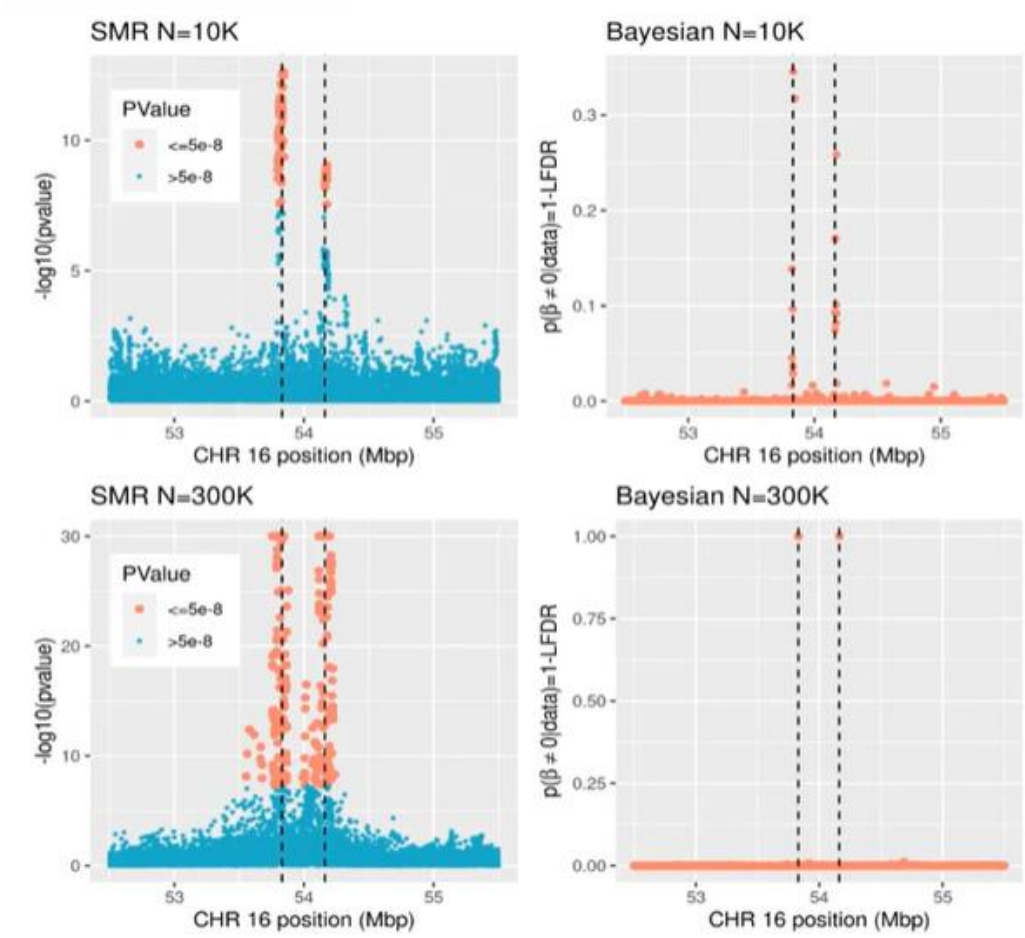

Figure 2. Performance of SMR and BVS ("Bayesian") methods with sample sizes of 10k and 300k. Plots display distributions of metrics of association across the DNA sequence (c16.52Mbp-56Mbp). Dotted lines indicate the positions in the DNA sequence of the 2 simulated causal SNPs. Adapted from [4].

Using the naïve SMR approach and a sample size of 10,000, he was able to locate the positions of the causal SNPs with relatively high SP. However, when increasing the sample size to 300,000, he found that the model predicted locations on the chromosome with lower SP. This kind of issue, where increasing sample size leads to less specific predictions (i.e. higher FDR), is

common with simple statistical methods like SMR because they cannot accommodate for the inherent noise that is associated with a larger sample size [6].

On the other hand, he was able to make predictions that were more specific than SMR, in a way that FDR decreases with increasing sample size, utilizing the more complex BVS method [6]. This highlights a major advantage of methods with more complexity, where predictions are not only more specific to begin with but increase as more observational data is gathered. Note that an advantage should be expected in the context of BVS, as existing knowledge is incorporated into the model before fitting, and this **prior** is adjusted based on the data in the sample [7], whereas SMR attempts to fit associations without prior information. In neural networks (NNs), such existing knowledge can be incorporated via **transfer learning** procedures [18].

Modern ML in Epidemiology

In her talk, *A Look into the Black Box: Machine Learning Applications in Perinatal Epidemiology*, Julie Petersen, PhD, presented her work on the utilization of ML in perinatal epidemiology. She explains that the overall goal of epidemiology is to identify modifiable agents to adverse health outcomes, but that this can be challenging because of the reliance on observational data, used because randomized controlled trials (RCTs) are often unethical, infeasible, and/or too resource intensive. However, modern ML is particularly well-suited to the field of epidemiology as it can uncover associations in large, complex data sets that may not be apparent using traditional statistical methods like SMR and BVS, potentially allowing for a more nuanced understanding of the relationships between risk factors and disease dynamics [8].

Supervised ML encompasses methods from logistic regression to modern NNs. In summary, a supervised ML model is a statistical model designed to predict or classify an outcome (aka. output, response) based on a set of predictors (aka. inputs, features, confounders). Essentially,

it “learns” a procedure to map the input(s) to the outcome(s) [8]. In this sense, “learning” refers to minimization of the loss/error function, and it may involve simple procedures like ordinary least squares (OLS) for linear regressions, or complicated procedures based on the principles of backpropagation and stochastic gradient descent (SGD) for NNs [9]. The technical details of these concepts are outside the scope of this paper, and they can be effectively thought of as mathematical procedures for determining model parameters that result in the lowest collective loss/error over the training data (i.e. the best fit).

Petersen attempted to identify risk factors for gastroschisis, a congenital malformation (birth defect), by applying supervised ML, specifically a **tree-based method** called *random forest*. This method works by randomly assigning combinations of predictors to the nodes (5 per node in this study) of many decision trees (specifically 1,000 in this study). These models are predictive, as the decision tree can be applied to a data point with a certain set of predictor values to output probability of the outcome of interest, given the path taken in the decision tree [8,10]. A visual depiction of a simple tree, with 1 variable per node, is shown in Figure 3.

Based on AUC calculations for the decision trees generated, Petersen was able to systematically determine the 10 most important predictors of gastroschisis out of a selected subset of 50. She subsequently ran logistic regressions incorporating these predictors to calculate odds ratios and p values, ultimately finding that maternal age and BMI were significantly associated with gastroschisis [8].

Unsupervised ML attempts to identify patterns within data, without explicitly defining outcomes, as the “outcomes” are learned by the model itself. In essence, the goal of this approach is to “learn” a set of subgroups (“outcomes”), each with distinct characteristics defined according to differences in the predictors.

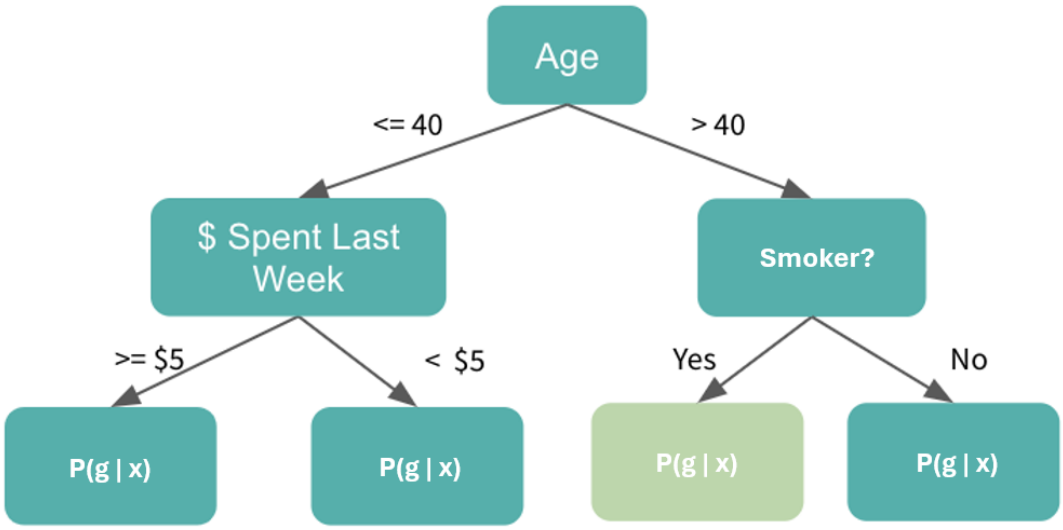

Figure 3. Example of a simple decision tree with 1 feature per node. There are 7 total **nodes** in this tree. The **root node** is "age." **Internal nodes** include "$ spent last week" and "Smoker?" The **terminal nodes** represent the outcome, $P(g|x)$, the probability of having gastroschisis given the path traveled on the tree. Adapted from [10].

For example, Petersen tried to identify whether placental abnormalities are associated with adverse outcomes of pregnancy, as suggested by previous literature but never reliably proven because there is no meaningful clinical classification procedure for placental abnormalities. She did this by first employing an unsupervised ML method to generate clusters for a set of more than 50 macroscopic and microscopic placental features, based on a sample of 2,005 distinct placentas. Doing so, she was effectively able to perform **dimensionality reduction**, *embedding* the original (50+)-dimensional data as 2-dimensional and 5-dimensional data. In other words, Petersen found that, based the 50+ selected placental features, the model "learned" 2 main types (2 "dimensions"/"clusters") of placental abnormalities, which she labeled "inflammation" and "vascular malperfusion" based on manual analysis of the features associated with each type. Moreover, it "learned" that these main types can be further subclassified into 5 sub-types (5 "dimensions"/"clusters"), labeled by Petersen as: "type 1 inflammation," "type 2 inflammation," "fetal malperfusion," "severe maternal malperfusion," and "mild maternal malperfusion" [8].

Relative risks of distinct adverse outcomes of pregnancy were calculated based on the 5 sub-types of placental abnormalities, and selected results of this analysis are shown in Table 2. She found that the "severe maternal vascular malperfusion" (SMVM) sub-type is significantly associated with preterm birth and that the "type 2 inflammation" sub-type is significantly associated with small head circumference (HC), but that neither of these sub-classifications are significantly associated with preeclampsia [8]. These results indicate that the model "learned" how to classify placental abnormalities in a way that is associated with distinct adverse outcomes, without any prior knowledge of these outcomes, providing further evidence (on top of the plausible groupings of placental abnormalities) that the model learned a clinically meaningful way to classify placental abnormalities. Moreover, this study provides evidence that placental abnormalities are not associated with preeclampsia.

Table 2. Crude relative risks (with 95% CIs) of 3 selected adverse pregnancy outcomes given placentas in the "normal," "SMVM," or "type 2 inflammation" clusters predicted by the unsupervised ML method. Bolded values indicate statistically significant findings. Adapted from [8].

	Normal	SMVM	Inflammation 2
Preterm birth	1.0	**1.9 (1.3, 2.9)**	0.7 (0.4, 1.3)
Small HC	1.0	2.3 (1.0, 5.0)	**2.5 (1.3, 4.7)**
Preeclampsia	1.0	1.2 (0.5, 3.1)	0.8 (0.3, 2.0)

Overall, Petersen's studies collectively demonstrate how both supervised and unsupervised ML methods can be incorporated alongside established epidemiological practices to discover patterns in data that might be difficult to discern otherwise.

Modern ML in Medicine

As touched on previously, CNNs are particularly well-suited for "learning" complex associations between predictors and outcomes, especially when those predictors are images.

Therefore, medical researchers have applied CNNs to a wide range of medical imaging modalities, including images of histopathological specimens. In *Digital pathology enhanced prostate cancer recurrence risk prediction in African Americans*, Albert Levin, PhD, the Director of the Center for Bioinformatics at Henry Ford Health Center, introduces a CNN that can effectively classify prostate cancer [15].

The ultimate purpose of a biopsy is to collect a histopathological specimen (HS) that can later be analyzed by a pathologist to determine whether the tissue is malignant [15]. This procedure can be roughly summarized as:

$$\text{Pathologist}(\text{HS}) = P(\text{Cancer} \mid \text{HS})$$

indicating that the HS is interpreted by a pathologist to predict the likelihood of cancer, given the features of that HS. In the context of a CNN, this is summarized as:

$$\text{CNN}(\text{HS}) = P(\text{Cancer} \mid \text{HS})$$

indicating that the HS is interpreted by a CNN to predict the likelihood of cancer, given the features in that HS.

Prostate cancer accounts for about 14% of diagnosed malignancies, while only accounting for 4.8% of all cancer-related deaths, indicating that most men with prostate cancer die with it, not because of it. Therefore, patients with mild prostate cancer may appreciate a greater benefit from less treatment, as aggressive chemotherapy involves many adverse effects [15,16].

Considering this, the Gleason Grade Group was developed to predict prognoses of prostate cancer, which aids in treatment decisions. This system classifies prostate cancers from Grade 1, the mildest form, to Grade 5, the most severe form, based on specific tissue types found on examination of HS. Treatment decisions for Groups 1, 4, and 5 are relatively straightforward;

however, there is a lack of clarity for treatment of Groups 2 and 3, the intermediate forms of prostate cancer [15].

Recent guidelines have recommended that pathologists report the percentage of a particular tissue type, called Gleason pattern 4 (GP4), when examining a HS in the context of prostate cancer because this value has been shown to have a high AUC in the prediction of prostate cancer prognosis. However, accurate assessment of this percentage is challenging for human uropathologists, and even more challenging for unspecialized pathologists, which prevents this recommendation from being widely implemented in practice [15].

Thus, Levin trained a CNN that predicts the probability that a particular region in a HS, HS_i, exhibits GP4, based on data labeled by a practicing uropathologist [15]. This model can be summarized by the equation:

$$\mathrm{CNN}(\mathrm{HS}_i) = P(\mathrm{GP4} \mid \mathrm{HS}_i)$$

Levin found that this model was highly effective at reproducing the GP4 gradings that uropathologists gave, as demonstrated in Figure 4. Further, he was able to apply this model across entire histopathological sections to estimate the percentage of GP4, %GP4, according to:

$$\%\mathrm{GP4} = \frac{1}{N}\sum_{i}^{N} \mathrm{CNN}(\mathrm{HS}_i)$$

which represents taking the mean of model predictions across all N regions of the entire HS.

Of note, Levin emphasized the importance of the interdisciplinary team that he is a part of, alongside the limitations that come with using 1 uropathologist to label the data, including poor generalizability and limited training data [15]. However, his results suggest that large NNs may be able to effectively apply collective decision making of consensus boards to guide individual providers, particularly in the context of highly specialized fields like uropathology.

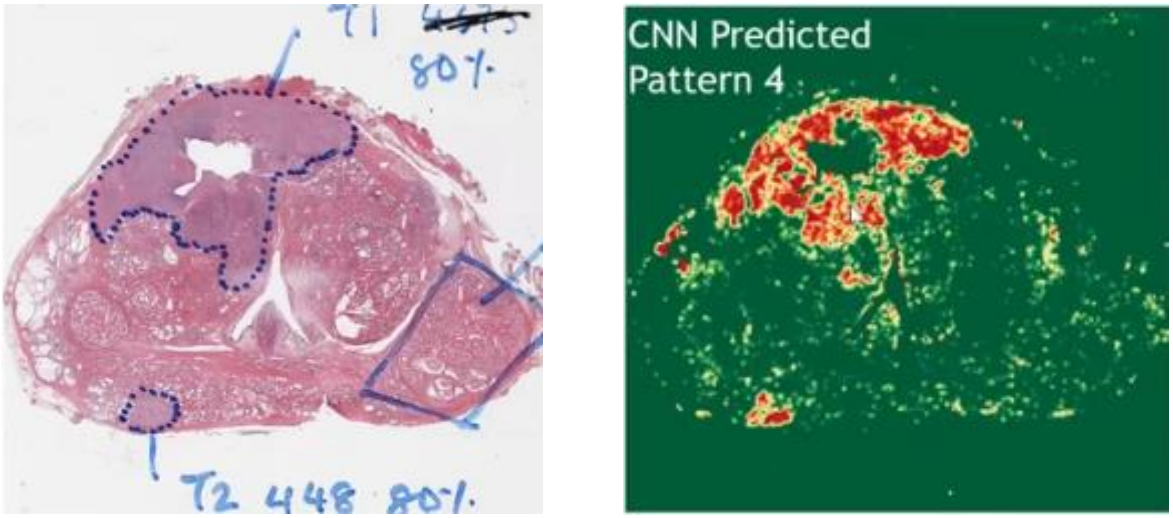

Figure 4. Side-by-side comparison of uropathologist (left) and CNN (right) detection of GP4. The left image is an H&E-stained prostate tissue section, where regions labeled T1 and T2 indicate suspected GP4 per the uropathologist. The right image illustrates CNN predictions of GP4 for the exact same tissue, where red indicates a higher probability. Adapted from [15].

Causal Estimation with ML

In *Causal Estimation of Exposure Shifts with Neural Networks: Evaluating the Health Benefits of Stricter Air Quality Standards in the US*, Mauricio Tec, PhD, a research associate at the Harvard School of Public Health, discusses a significant advancement in establishing causality with ML in the field of epidemiology [11].

Shift-response functions (SRFs) are important for understanding how changes in exposure, like those to air pollutants, affect health outcomes. Put simply, they can be used to model how a particular change in exposure affects a selected health outcome. The type of change in exposure determines the type of SRF; for example, Figure 5 shows 2 common types of SRFs, percent-reduction exposure shifts and cutoff exposure shifts, which correspond to policy that reduces air pollution by 10% everywhere and policy that implements a cutoff for the maximum amount of air pollution allowed in any location, respectively [11].

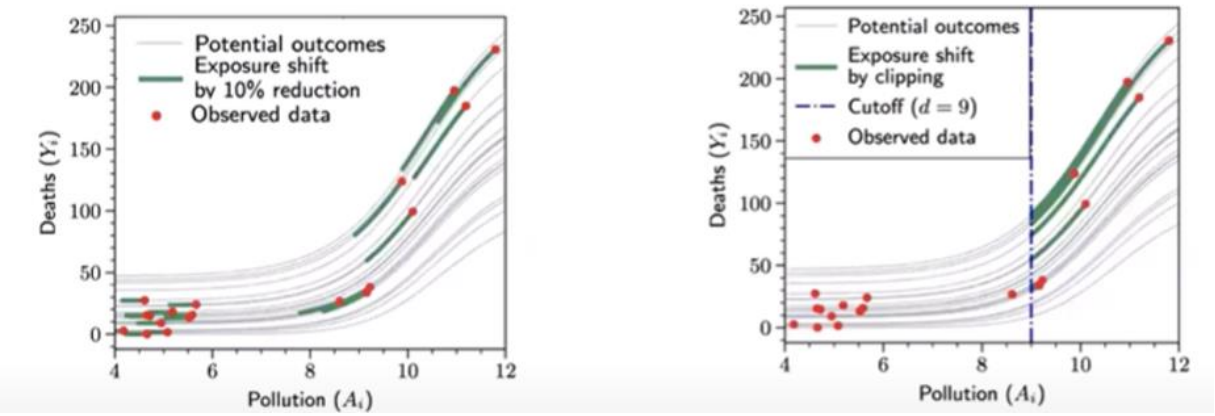

Figure 5. Visualization of applying SRFs to a dataset. Green lines indicate the shifts calculated for each data point, based on the SRF. Plots show death toll versus level of pollution, and points represent individual ZIP-codes in the US. The plot on the left shows a percent-reduction exposure shift and the plot on the right shows a cutoff exposure shift. Adapted from [11].

Tec used SRFs to model potential regulations that could be incorporated into the US National Ambient Air Quality Standards (NAAQS). Doing so, he wanted to develop a ML model to predict these SRFs, to determine whether such regulations have a causal effect on improving health outcomes [11]. He explains that predicting causation with ML can be achieved by incorporating a framework called Targeted Regularization (TR), which involves 5 main steps, outlined on the next page [11].

By incorporating this framework, Tec was able to develop a robust ML model for predicting SRFs in the context of NAAQS [14]. He specifically evaluated whether revising NAAQS for fine particles from 12 $\mu g/m^3$ to 9 $\mu g/m^3$ would have a significant impact on health. By training TRESNET to predict a SRF, specifically a cutoff exposure shift, he estimated that this policy would lead to a 4% total reduction in death across the US [11]. Of note, Tec's TR framework can generally be thought of as a way to general outline for developing modern ML models. By implementing the underlying principles of each step (outlined on the next page) into a ML project, researchers can ensure their models are implemented responsibly and are of high quality.

1. **Perturb the model.** Introduce slight variations (i.e. random noise) into the predictor data. This prevents the ML model from "memorizing" spurious correlations in the specific data you provide the model [11].
2. **Choose loss function.** As discussed previously, ML training relies on minimization of a loss function, and choosing a loss function that is suitable to a specific application is crucial. For example, in Tec's study, a loss function that causes underrepresented data points to have more weight in the overall loss was utilized to account for extremes in the data (e.g. exceptionally high pollution rates). If not incorporated, the model may "learn" a mapping that is useful for most data points, but not for extreme cases [11].
3. **Choose NN architecture.** In ML, choosing a NN architecture suitable for your data is key. For instance, CNNs were mentioned earlier in the context of visual echocardiogram data as they have proven effective in image recognition applications [12]. In Tec's study, he designed an architecture called TRESNET, which was designed based on the existing DragonNet, as it has been proven effective for the simultaneous estimation of treatment effects and outcome predictions, a goal of his study [11,13].
4. **Validate with simulation studies.** To determine whether an ML model can predict causation, it should be evaluated with data that is simulated under controlled conditions. By validating a model against such data, researchers can demonstrate reliability before applying it to real-world data.
5. **Establish Theoretical Guarantees.** This step aims to clarify the applications and limits of the model, which is integral to responsible research and innovation. In Tec's study, mathematical proof was provided demonstrating that TRESNET was indeed doubly robust [11] and the detailed proof is included in [14], but this is outside the scope of this paper.

Concluding Remarks

There are several limitations in biostatistical ML including the frequently discussed issues of limited training data and low explainability (i.e. the "black box" of ML). Seemingly not as frequently mentioned is the potential for unexpected consequences with the application of this powerful technology in a sensitive field like healthcare. For example, statistical algorithms that are already applied on roughly 200 million American patients annually have been shown to reproduce racial bias, via societal patterns reflected in training data and inherent biases of physicians and researchers [17].

Despite these limitations, which can often be overlooked amidst the current AI hype, modern ML holds significant promise. As shown in this project, it has several potential applications in epidemiology and medical diagnostics, typically in conjunction with traditional methods but now emerging in standalone use cases. Moreover, there are significant efforts being made to address the challenges associated with ML: simulations, like those in studies outlined in this project, are being used to augment limited training data sets; transfer learning is being utilized to enhance model robustness [18]; developments in explainable AI (XAI) aim to make predictions of ML models more understandable to humans [19]; and there has been increasing focus on a responsible approach to research and innovation [20], as exemplified in the deliberate process of TR described by Mauricio Tec [14]. These efforts may help researchers overcome the technical barriers of biostatistical ML, while also mitigating the ethical, social, and environmental risks associated.

Overall, with so many promising results and ongoing developments, the future of biostatistical ML should be quite exciting, and it will be interesting to see what advancements are on the horizon.

References

1. Florkowski C. M. (2008). Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *The Clinical biochemist. Reviews, 29 Suppl 1*(Suppl 1), S83–S87. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2556590/
2. Ellmers, L. J., Knowles, J. W., Kim, H. S., Smithies, O., Maeda, N., & Cameron, V. A. (2002). Ventricular expression of natriuretic peptides in Npr1(-/-) mice with cardiac hypertrophy and fibrosis. *American journal of physiology. Heart and circulatory physiology, 283*(2), H707–H714. https://doi.org/10.1152/ajpheart.00677.2001
3. Heidenreich, P. A., Bozkurt, B., Aguilar, D., Allen, L. A., Byun, J. J., Colvin, M. M., Deswal, A., Drazner, M. H., Dunlay, S. M., Evers, L. R., Fang, J. C., Fedson, S. E., Fonarow, G. C., Hayek, S. S., Hernandez, A. F., Khazanie, P., Kittleson, M. M., Lee, C. S., Link, M. S., Milano, C. A., … Yancy, C. W. (2022). 2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation, 145*(18), e895–e1032. https://doi.org/10.1161/CIR.0000000000001063
4. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression (3rd ed.)*. http://onlinelibrary.wiley.com/book/10.1002/9781118548387
5. Krittanawong, C., Omar, A. M. S., Narula, S., Sengupta, P. P., Glicksberg, B. S., Narula, J., & Argulian, E. (2023). Deep Learning for Echocardiography: Introduction for Clinicians and Future Vision: State-of-the-Art Review. *Life (Basel, Switzerland), 13*(4), 1029. https://doi.org/10.3390/life13041029
6. De los Campos, Gustavo. (2021, October 14). *Powerful & Safe: Using Bayesian Variable Selection Models to Map Risk Variants with Biobank-size Data*. [Lecture]. Department of Epidemiology and Biostatistics at Michigan State University, East Lansing, MI, USA. https://youtu.be/zeZiJO9yXfU
7. Robert, Christian. (1994). From Prior Information to Prior Distributions. *The Bayesian Choice. New York: Springer*. pp. 89–136. ISBN 0-387-94296-3.
8. Peterson, Julie. (2022, October 13). *A Look into the Black Box: Machine Learning Applications in Perinatal Epidemiology*. [Lecture]. Department of Epidemiology at the University of Pittsburgh, Pittsburgh, PA, USA. https://youtu.be/-iFjE9yrb8k
9. Amari, Shunichi. (1992, September 16). Backpropagation and stochastic gradient descent method. *Neurocomputing 5*, 184-196. https://bsi-ni.brain.riken.jp/database/file/141/142.pdf
10. Gross, Katie. (2020). *Tree-Based Models: How They Work*. https://blog.dataiku.com/tree-based-models-how-they-work-in-plain-english
11. Tec, Mauricio. (2024, February 2024). *Causal Estimation of Exposure Shifts with Neural Networks: Evaluating the Health Benefits of Stricter Air Quality Standards in the US*. [Lecture]. Department of Biostatistics at the Harvard School of Public Health, Boston, MA, USA. https://youtu.be/vxDa-qi73UY
12. Hijazi, S., Kumar, R. and Rowen, C. (2015) *Using Convolutional Neural Networks for Image Recognition*. 1-12

13. Shi, C., Blei, D. M., & Veitch, V. (2019, October 17). *Adapting neural networks for the estimation of treatment effects*. [Preprint]. https://arxiv.org/abs/1906.02120v2
14. Tec, M., Mudele, O., Josey, K., & Dominici, F. (2023, December 6). *Causal estimation of exposure shifts with neural networks: Evaluating the health benefits of stricter air quality standards in the US (Version 3)* [Preprint]. https://arxiv.org/abs/1906.02120v2
15. Levin, Albert. (2023, February 9). *Digital pathology enhanced prostate cancer recurrence risk prediction in African Americans*. [Lecture]. Center for Bioinformatics at Henry Ford Health Center, Detroit, MI, USA. https://youtu.be/U7SzQrdZssk
16. American Cancer Society. (2024). *Chemotherapy side effects.* https://www.cancer.org/cancer/managing-cancer/treatment-types/chemotherapy/chemotherapy-side-effects.html
17. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science (New York, N.Y.), 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342
18. Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., & Ganslandt, T. (2022). Transfer learning for medical image classification: a literature review. *BMC medical imaging*, *22*(1), 69. https://doi.org/10.1186/s12880-022-00793-7
19. Confalonieri, R., Coba, L., Wagner, B. & Besold, T. R. (2021). A historical perspective of explainable Artificial Intelligence. *WIREs Data Mining Knowl Discov 11*(1), https://doi.org/10.1002/widm.1391
20. Felt, U., Rayvon, F., Miller, C. A. & Smith-Doerr, L. (2016). Responsible Research and Innovation. in *The Handbook of Science and Technology Studies, fourth edition (The MIT Press) (ed. Moore, K.)*. 853–881.