Pulling the Plug: Equitable Guidelines for Machine Learning Neuroprognostication

Abstract

Neuroprognostication is the field of predicting recovery from comas or other disorders of consciousness after experiencing brain injury. Unfortunately, many comatose patients are withdrawn from life-sustaining treatment by medical professionals if they are predicted to have a poor outcome within a certain timeframe. This practice is a confounding factor in clinical neuroprognostication studies because of the human bias known as self-fulfilling prophecy, whereby taking a patient off life support because of a prediction turns out to be the reason they die when they may have lived otherwise. In recent years, the growth of machine learning has resulted in the creation of coma prognostication algorithms in order to improve patient care and healthcare decision-making. This paper proposes and elaborates on a data equity-informed approach to creating equitable guidelines for machine learning prognostication by concentrating on the data lifecycle and medical staff training.

Pulling the Plug: Equitable Guidelines for Machine Learning Neuroprognostication

Introduction

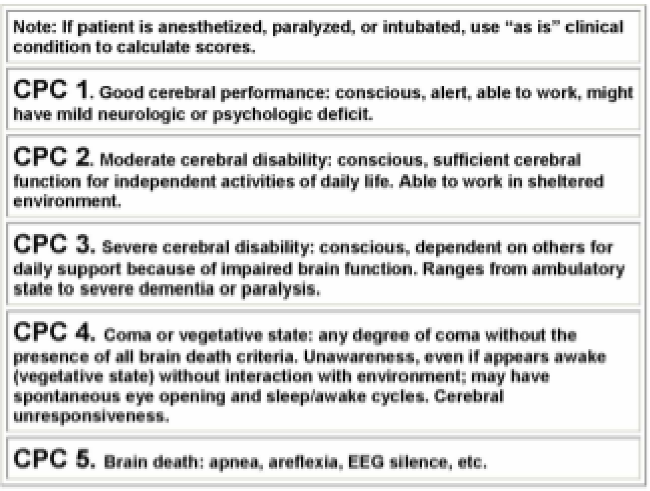
Archie Battersbee, a 12-year old boy, died in a London hospital on August 5, 2022. After remaining in a coma for four months, his parents begged doctors not to withdraw life-sustaining treatment such as his medications and ventilator because they were convinced he would wake up. This plea was rejected in Archie's best interests since the hospital argued he had "no chance of recovery," (Kirka).

Coma prognostication is one of the most difficult areas in the medical field because of the level of uncertainty and unpredictability of recovery estimates. While ascertaining the numerical estimates of coma incidence has not been feasible in the past, one research group found that there are at least 250 annual coma cases for every 100,000 people (Kondziella et al.). That amounts to more than 20 million people affected globally. Many of these individuals will not wake up unless they experience adequate brain activity within a crucial time frame. Gorelova, a University of Pittsburgh corresponder, comments, "It often takes two weeks for [traumatic brain injury (TBI)] patients to emerge from their coma and begin their recoveries—yet severe TBI patients are often taken off life support within the first 72 hours after hospital admission," (Gorelova and Davis).

In order to enhance patient care and outcomes, predictive machine learning (ML) has been gaining traction in prognosis efforts. In the context of comatose patients, this subset of ML refers to the creation of systems by scholars that predict the prognosis, or likely outcome, of a coma by training algorithms on information, known as a dataset. One limitation to these algorithms is the many biases that result from their creation, from overreliance on ML tools, known as automation bias, to data bias, where aspects of data are over- or under-represented. This raises the question, To what extent do withdrawal of life-sustaining therapy and self-fulfilling prophecy play a role in coma prognostication algorithms in healthcare, and how can this inform future guidelines and decision-making?

History and Current Machine Learning Approaches

In the past few years, researchers have started creating ML methods to predict neurological outcomes in comatose patients. Many of these models utilize fMRI or EEG data to train and test a model. These models analyze the data and learn patterns to predict a good or poor outcome. For example, Gorelova's sentiment about withdrawal of life support was a response contrasting traditional approaches with a predictive algorithm created by Shandong Wu, an associate professor of radiology and bioengineering at the university. The algorithm used brain imaging data to predict 6-month survival and recovery rates of patients, which would inform doctors about when to withdraw treatment (Pease et al.). Algorithms further add context by using well-known objective scales in the world of coma prognostication, including the Glasgow Coma Scale and Cerebral Performance Category (CPC) Scale to make results more interpretable. Figure 1 has descriptions of each cerebral performance category, with 1-2 predicting a good outcome and 3-5 predicting a poor outcome.



Note: If patient is anesthetized, paralyzed, or intubated, use "as is" clinical condition to calculate scores.

CPC 1. Good cerebral performance: conscious, alert, able to work, might have mild neurologic or psychologic deficit.

CPC 2. Moderate cerebral disability: conscious, sufficient cerebral function for independent activities of daily life. Able to work in sheltered environment.

CPC 3. Severe cerebral disability: conscious, dependent on others for daily support because of impaired brain function. Ranges from ambulatory state to severe dementia or paralysis.

CPC 4. Coma or vegetative state: any degree of coma without the presence of all brain death criteria. Unawareness, even if appears awake (vegetative state) without interaction with environment; may have spontaneous eye opening and sleep/awake cycles. Cerebral unresponsiveness.

CPC 5. Brain death: apnea, areflexia, EEG silence, etc.

Figure 1: Cerebral Performance Category (CPC) Scale Descriptions (Safar)

Many studies predict coma recovery outcomes based on datasets made up of electroencephalogram (EEG) data, which are recordings of brain activity. The model trains on thousands of pieces of data just like the sample below in Figure 2 in order to detect patterns in the data and create an accurate model.

The benefit of ML tools is that they can be used in conjunction with professionals as a second opinion of sorts. In 2018, a system predicting recovery scores for comatose patients proved this sentiment true. A group of doctors in China assessed seven patients with very low recovery scores. The system gave them close to full scores and a predicted recovery of within a year (Chen). The patients all recovered. In a world without ML, the hospital would likely have taken the patients off life support.

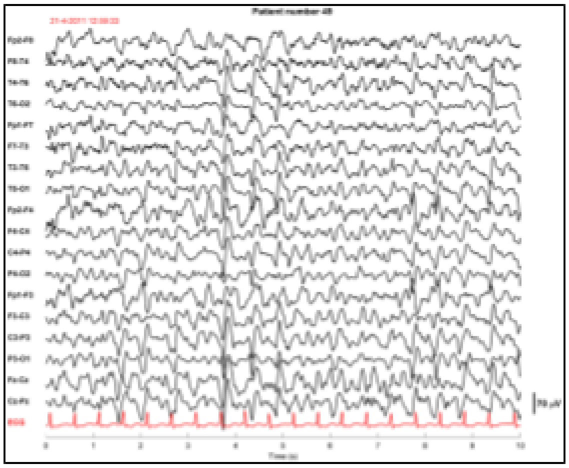


Figure 2: Sample EEG Data (Hofmeijer)

Self-fulfilling Prophecy

Consider Archie's accident. On April 6, Archie's mom found him unresponsive after a cardiac arrest due to strangulation. He was rushed to the hospital, resuscitated, and confirmed to have suffered hypoxic ischemic brain injury (caused by lack of oxygen to the brain). For the next four months, Archie was monitored in his coma, and traditional prognostication techniques were probably used to determine brain damage and recovery potential. The damage was severe and prognosis was poor, so medical professionals suggested withdrawal of life-sustaining treatment (WLST).

In cases like Archie's, the age-old adage "To expect defeat is nine-tenths of defeat itself," has the potential to take on a much graver meaning. Who decides when to "pull the plug"? How can medical professionals be sure they aren't making an incorrect evaluation? In some tragic cases, a patient could have lived if the doctor decided not to withdraw treatment. Can this horrifying reality be prevented? It all comes down to medical decision-making and the human bias known as a self-fulfilling prophecy (SFP). SFP in neuroprognostication occurs when “a patient in coma is predicted to have a poor outcome, and life-sustaining treatment is withdrawn on the basis of that prediction, thus directly bringing about a poor outcome (viz. death) for that patient,” (Mertens et al.).

SFP and Machine Learning

SFPs have long been studied in fields that use clinical trials and pose a concern to healthcare professionals. However, their significance to predicting clinical outcomes with ML is just being realized due to the recent emergence of these technologies. As a result, the interplay between the two concepts in coma recovery is understudied, and there is not an abundance of literature published. Existing exploration of the topic suggests that ML can both mitigate or exacerbate SFPs.

As previously mentioned, the system created by the Chinese Academy of Sciences saved seven patients. The scientists even claimed, “The possible prediction of the recovery of patient consciousness will directly affect the choice of clinical treatment strategies, and even the choice of life or death by the patient’s relatives,” (Chen). This suggests that such models can mitigate SFP and prevent countless deaths.

On the other hand, these algorithms can amplify existing human bias. Maria De-Arteaga, a researcher working in algorithmic fairness and human-AI interaction, is one of the first to acknowledge the challenge posed by SFP to ML by diving into this theme through a scoping review. De-Artega concludes, “Models (and providers) trained to predict outcomes based on data available prior to transfer may learn erroneous relationships between clinical patterns that predict” (De-Artega and Elmer) the decision to withdraw treatment rather than the poor outcome itself. This is how medical professionals' SFP is encoded into data and therefore model predictions.

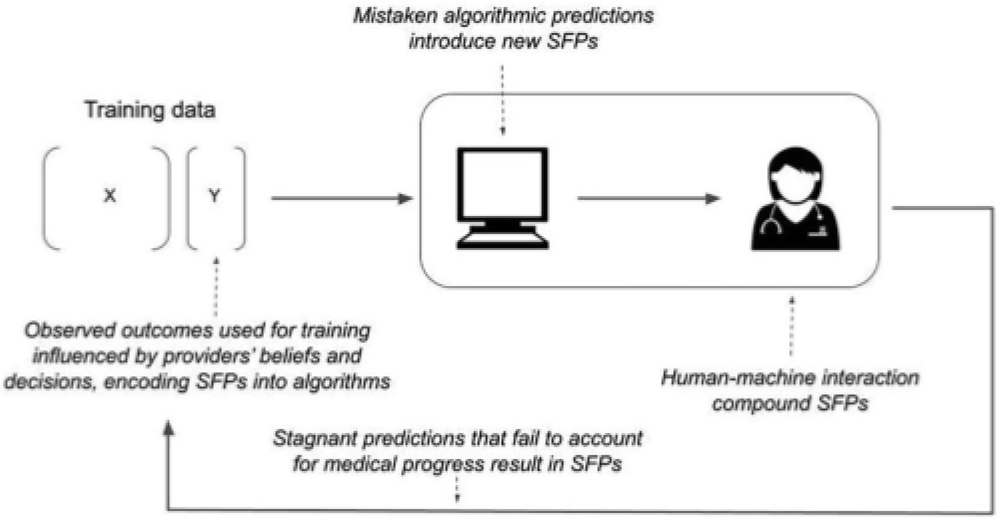

Mistaken algorithmic predictions introduce new SFPs
Training data
X
Y
Observed outcomes used for training influenced by providers' beliefs and decisions, encoding SFPs into algorithms
Human-machine interaction compound SFPs
Stagnant predictions that fail to account for medical progress result in SFPs

Figure 3: SFPs in Machine Learning (De-Arteaga and Elmer)

The term "feedback loop" is often used when discussing SFP in psychology. A feedback loop is "a system in which two or more aspects of the system influence each other," (Loper). Figure 3 outlines De-Artega's findings with the dangerous feedback loop created by algorithms and humans. It details how human-machine interactions compound SFPs by encoding SFPs into algorithms with flawed data, which in turn creates more SFPs in clinical settings and repeats the cycle. This means if there is a data point where a doctor wrongly suggested WLST for a Patient X — leading to Patient X's unnecessary death — then this data point could have long-lasting consequences on the algorithm and its implementation.

A metric that has showcased this amplification already exists. It is called the false positive rate (FPR) of an algorithm and is increased when the algorithm predicts a poor outcome falsely, when it should actually be a good outcome. One study looked at the FPR of the absence of one biomarker of poor outcome, called somatosensory evoked potential (SSEP). The authors defined FPR in this case as "the proportion of subjects with initially absent SSEPs who would eventually achieve a good neurological recovery in a setting where life-sustaining therapy is continued indefinitely," (Amorim et al.). They showed that the general FPR estimate of SSEP was around .7%, but after accounting for the average rate of WLST in the studies they reviewed, the calculated FPR came out to be 7.7%. This is 11 times higher than what was widely accepted by the general public at the time, which demonstrates the amplification effect of SFPs. This study of the SSEP biomarker was applicable to traditional clinical prognostication, but this reasoning can be extended to ML where computers extract patterns from data since they also require similar consideration of WLST. Lack of such consideration could lead to a much higher FPR in algorithms.

Data Guidelines

In order to account for the aforementioned factors within the feedback loop pointed out in De-Arteaga's study, it is clear that the training data of such models must be addressed. In fact, De-Arteaga claims, "Training models on data accrued in settings [where] withdrawal of life-sustaining therapies is prohibited or strictly protocolized and potential confounders are minimized can help prevent providers' biases, mistaken beliefs and clinical choices from becoming encoded in algorithmic predictions," (De-Artega and Elmer). Tracking the source of training data is imperative to create comprehensive models for coma recovery prediction. At the crux of the solution lies data equity. Data equity is "the consideration, through an equity lens, of the ways in which data is collected, analyzed, interpreted, and distributed," (Lee-Ibarra) and it must be present throughout the data lifecycle. Parallels can be drawn between the feedback loop and these four target areas since they are the main stages of the data lifecycle (Gaddy and Scott) according to correspondents from the Urban Institute, an organization striving to advance equity through research. One caveat of SFP is that it is rooted in human decision-making and thoughts.

This is why it is necessary to bring the concept of data equity into conversation with WLST to understand how bias can prevent equity in the neuroprognostication field and the guidelines needed for a resolution.

Equitable data collection methods should be the standard in neuroprognostication research, which raises the concern of whether this is ensured when it comes to data inputs for such algorithms (See "History and Current Machine Learning Approaches"). "Systemic Racism in EEG Research: Considerations and Potential Solutions," is a theoretical paper which dives into some of the rampant racial bias in the EEG research field over time and the exclusion of marginalized communities in EEG datasets (Choy et al.). This exclusion has been tied to Black hairstyles and hair texture, which complicates the attachment of electrodes to collect data. In a recent attempt to change the course of EEG research, a group of undergraduates created Sevo electrodes, which harness the transformation of a traditional African hairstyle. The electrodes are fashioned as a hair clip in order to separate coarse and voluminous hair and successfully attach to the scalp (Etienne et al.). As EEGs are one of the most common data inputs in neuroprognostication algorithms, preventing exclusionary traditions is a precursor to dealing with WLST in data.

Data bias stemming from WLST can be broken down into subcategories. A Duke study analyzing how brain injury patients are treated showed that "race, geographic region, and payment status were significantly associated with the decision to withdraw life support," (Williamson et al.) as well as patients with Medicare in contrast to those with other means of insurance. Even the number of neurosurgeons present in the hospital carried immense weight when analyzing association with WLST. This is why neuroprognostication algorithms must take into account other biases relevant to SFP in order to minimize SFP and be truly reliable for patients.

While neuroprognostication algorithms have not risen to this challenge yet, biases in healthcare algorithm data usage have begun to be addressed in different clinical prediction settings. For instance, in order to reduce bias when predicting the occurrence of postpartum depression, authors from the IBM TJ Watson Research Center incorporated two debiasing approaches in their algorithm. The first was preprocessing of data, which means working to remove the bias before data analysis, by re-weighting based on bias. The second was in-processing by minimizing prejudice while training on the dataset (Park et al.). When working with vast amounts of data for training these algorithms, preventing human bias during data collection may not always be feasible. Debiasing mandates in algorithm design would be an alternative that seeks to prevent encoding SFPs into algorithms.

Training Guidelines

After addressing data-driven concerns, the regulation of researchers and medical professionals still remains. Mayli Mertens, an associate researcher at Copenhagen University, states in her staff recommendations in the *Journal of Medical Ethics*, “When possible, treating medical staff should be completely blinded from neuroprognostic studies,” (Mertens et al.) so that their decision is not influenced by the prediction of the algorithm, which can further introduce SFP. This means they should not be aware of the algorithm or technological tool being tested. This is similar to De-Artega’s conclusions about how human-computer interactions can amplify SFP in data. Blinding is a well-known technique applied in clinical research to prevent bias not limited to SFP. There has been no research into how unblinded studies can interact with bias in the context of ML studies. Consequently, unblinded studies could encode bias into algorithms. In coming years, algorithmic studies may begin to focus on clinical testing, especially with the exponential growth of artificial intelligence. Thus, blinding should be a prerequisite for conducting algorithmic clinical practice research.

Finally, guidelines would not be complete without thinking about how these algorithms will eventually be implemented in a healthcare setting. After tackling equity in data collection and analysis, the final stages are interpretation and distribution. The overlap of research and implementation is extremely important in the field of medicine, and one area where this comes into play is in the evaluation of new tools. This overlap relies on a key building block in ML approaches for neuroprognostication: interpretability. Medical interpretability refers to "a degree to which a human can understand the cause of a decision from an ML model," (Abdullah et al.). If the doctor told Archie's parents he had a 95% chance of a poor outcome, would they really have a better comprehension of the situation? Conventional ML techniques operate using "black box" logic, which is less interpretable. The previous example of such logic is tied to the traditional binary classification model, which outputs a decimal value from 0 to 1 representing the sliding scale from a poor outcome to a good outcome. However, there are many factors indicative of a poor outcome, and interpretable algorithms grant more transparency and information to families of coma victims.

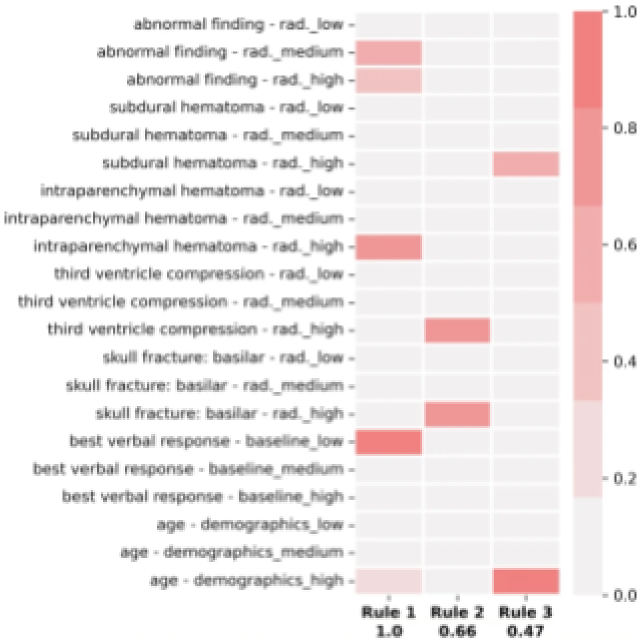


Figure 4: Interpretable rules extracted from patient data (Minoccheri et al.)

Interpretable algorithms try to employ “fuzzy logic” by integrating human concepts to make more sense of uncertainty. One visually exceptional example of this was conducted by researchers from the University of Michigan. The model predicted poor outcomes of patients hospitalized with traumatic brain injury. The cherry on top was that the algorithm achieved this goal by extracting “simple, human-understandable rules that explain the model’s predictions” (Minoccheri et al.), which can be seen above in Figure 4. The presence or absence of certain data can satisfy the rule and produce the likelihood of that rule at the bottom of the figure. Rather than a single number, these rules are clinically acceptable so medical professionals could use them as a tool to explain their final decision. This paper was only published in August 2022, so a deeper evaluation of interpretability standards of algorithms is necessary to aid medical professionals in developing more interpretable ML models.

Solution

Based on the above analysis, the proposed equitable guidelines for preventing bias in coma recovery algorithms should be rooted in minimizing human bias from WLST and maximizing data equity. This requires attention in two areas: the data lifecycle and medical staff training. Many other sources have alluded to the fact that WLST should be emphasized more with respect to data usage and professional decision-making. After a scoping search on Google Scholar and PubMed, the aforementioned opinions of De-Arteaga and Mertens were the only two sources with concrete guidelines-informed approaches to preventing SFP in ML for neuroprognostication. Existing standards like the “Standards for Studies of Neurological Prognostication in Comatose Survivors of Cardiac Arrest” by the American Heart Association inform researchers that “establishing a strict protocol for WLST, complying with it consistently, and carefully describing the causes of death in all patients may help control for the self-fulfilling prophecy bias," (Geocadin et al.). Protocols for WLST can mitigate SFP, but the field still lacks wider enforcement of equitable data guidelines that can standardize the data lifecycle and eliminate bias.

The implementation of ML in clinical settings is still a relatively new playing field. Mertens's publication from this November is one of the first to start directly addressing the shortcomings of technological neuroprognostication methods in the context of augmented SFP bias. It is significant to note that most of the literature considered in this paper has been conducted over the past five years, with the majority having been published in 2022. Technological neuroprognostication still has a long way to go before systems can come to fruition in clinical settings. Guideline-recommended algorithms have already been developed, such as the four-step algorithm informed by the European Resuscitation Council and the European Society of Intensive Care Medicine guidelines. The algorithm did not have false positives and was able to identify 38.7% of patients with a poor outcome (Moseby-Knappe et al.). WLST was permitted in the data according to specific criteria. The researchers, who work with the Department of Clinical Sciences at Sweden's Lund University, acknowledged the inherent bias in this approach because of this, asserting that "influence from the self-fulfilling prophecy cannot be excluded," (Moseby-Knappe et al.).

In order for guideline-recommended algorithms to see a significant change, support in the form of expert collaboration is needed. One initiative launched by the Neurocritical Care Society to fill this deficit was The Curing Coma Campaign, which strives towards the mission of "idea generation, expert consensus, and strategic planning," (Mainali et al.). The campaign's neuroprognostication team identified five gaps in efforts, one of which was the lack of standard methods for decision-making in WLST. Although the formal protocolization of WLST may be a long time coming, the proposed data and training techniques can be applied to make an impact in the present. Furthermore, they can be applied to other clinical contexts that deal with SFPs and inherent human bias.

The bottom line is that guidelines themselves must adapt to encompass new technologies and their interaction with human biases. This solution entails the adoption of a data equity section in future iterations of neuroprognostication guidelines and forums. By diminishing human bias and promoting data equity, these two-pronged guidelines will make it easier on the loved ones of patients like Archie to accept medical professionals' decisions by shedding more light on the reasoning behind them. Guidelines bring algorithms one step closer to becoming prevalent in clinical settings and one step closer to saving countless patients like Archie.

Works Cited

Abdullah, Talal AA, et al. "A review of interpretable ml in healthcare: Taxonomy, applications, challenges, and future directions." *Symmetry* 13.12 (2021): 2439. https://www.mdpi.com/2073-8994/13/12/2439.

Amorim, Edilberto, et al. "Estimating the false positive rate of absent somatosensory evoked potentials in cardiac arrest prognostication." *Critical care medicine* 46.12 (2018): e1213. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6424571/.

Chen, Stephen. "Doctors Said the Coma Patients Would Never Wake. Ai Knew Better." *South China Morning Post*, 10 Sept. 2018, https://www.scmp.com/news/china/science/article/2163298/doctors-said-coma-patients-would-never-wake-ai-said-they-would.

Choy, Tricia, et al. "Systemic racism in EEG Research: considerations and potential solutions." *Affective Science* 3.1 (2022): 14-20. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9383002/.

De-Arteaga, Maria, and Jonathan Elmer. "Self-fulfilling prophecies and machine learning in resuscitation science." *Resuscitation* (2022). https://www.sciencedirect.com/science/article/pii/S0300957222006931#f0005.

Etienne, Arnelle, et al. "Novel electrodes for reliable EEG recordings on coarse and curly hair." *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020. https://pubmed.ncbi.nlm.nih.gov/33019375/.

Gaddy, Marcus, and Kassie Scott. "Principles for advancing equitable data practice." *Washington, DC: Urban Institute* (2020).

https://view.ckcest.cn/AllFiles/ZKBG/Pages/141/principles-for-advancing-equitable-data-practice.pdf.

Geocadin, Romergryko G., et al. "Standards for studies of neurological prognostication in comatose survivors of cardiac arrest: a scientific statement from the American Heart Association." *Circulation* 140.9 (2019): e517-e542. https://www.ahajournals.org/doi/full/10.1161/CIR.0000000000000702.

Gorelova, Anastasia, and Sheila Davis. “Machine-Learning Model to Help TBI Patients.” *University of Pittsburgh Neurosurgery*, UPMC News Bureau, 26 Apr. 2022, https://www.neurosurgery.pitt.edu/news/machine-learning-model-help-tbi-patients.

Hofmeijer, Jeannette, et al. "Unstandardized treatment of electroencephalographic status epilepticus does not improve outcome of comatose patients after cardiac arrest." *Frontiers in neurology* 5 (2014): 39. https://www.frontiersin.org/articles/10.3389/fneur.2014.00039/full.

Kirka, Danica. “Boy at Heart of UK Court Battle Dies after Life Support Ends.” *US News*, Associated Press, 6 Aug. 2022, https://www.usnews.com/news/world/articles/2022-08-06/mother-of-comatose-uk-boy-says-hospital-to-end-care-soon.

Kondziella, Daniel, et al. "Incidence and prevalence of coma in the UK and the USA." *Brain Communications* 4.5 (2022): fcac188. https://academic.oup.com/braincomms/article/4/5/fcac188/6673810#373019526.

Lee-Ibarra, Joyce. “Data Equity: What Is It, and Why Does It Matter? .” Hawai'i Data Collaborative, 2 July 2020, https://www.hawaiidata.org/news/2020/7/1/data-equity-what-is-it-and-why-does-it-matter .

Loper, Chris. Feedback Loops. Northwest Educational Services. 22 Sep. 2014. https://www.nwtutoring.com/2014/09/22/feedback-loops/.

Mainali, Shraddha, et al. "Proceedings of the Second Curing Coma Campaign NIH Symposium: Challenging the Future of Research for Coma and Disorders of Consciousness." *Neurocritical Care* (2022): 1-25. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9283342/.

Mertens, Mayli, et al. "Can we learn from hidden mistakes? Self-fulfilling prophecy and responsible neuroprognostic innovation." *Journal of Medical Ethics* 48.11 (2022): 922-928. https://jme.bmj.com/content/48/11/922#ref-3.

Minoccheri, Cristian, et al. "An interpretable neural network for outcome prediction in traumatic brain injury." *BMC Medical Informatics and Decision Making* 22.1 (2022): 1-9. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-022-01953-z#Sec7.

Moseby-Knappe, Marion, et al. "Performance of a guideline-recommended algorithm for prognostication of poor neurological outcome after cardiac arrest." *Intensive care medicine* 46.10 (2020): 1852-1862. https://link.springer.com/article/10.1007/s00134-020-06080-9.

Park, Yoonyoung, et al. "Comparison of methods to reduce bias from clinical prediction models of postpartum depression." *JAMA network open* 4.4 (2021): e213909-e213909. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2778568.

Pease, Matthew, et al. "Outcome prediction in patients with severe traumatic brain injury using deep learning from head CT scans." *Radiology* (2022). https://pubs.rsna.org/doi/pdf/10.1148/radiol.212181.

Safar, Peter. "Resuscitation after brain ischemia." *Brain failure and resuscitation* 155 (1981): 155-184.

Williamson, Theresa, et al. "Withdrawal of life-supporting treatment in severe traumatic brain injury." *JAMA surgery* 155.8 (2020): 723-731. https://jamanetwork.com/journals/jamasurgery/fullarticle/2767404.