Literature Review #2: “Understanding the bias in machine learning systems”

The next paper we will discuss is “Understanding the bias in machine learning systems for cardiovascular disease risk assessment” written by Suri et al. and published in the journal *Computers in Biology and Medicine*. Written in the 142nd volume of the journal released back in March of 2022 this article was the first of its kind in terms of delivering a meta-analysis of bias across many different papers published in the field of cardiovascular disease (CVD) risk assessment. A meta-analysis is when the results of different studies are compared to each other using various statistical methods to determine trends across the papers and identify attempts to quantify differences in outcomes between them based on their respective approaches. The main difference in approach for the papers analyzed in this article are whether or not machine learning algorithms in their diagnostic predictor for CVD. The paper explains how machine learning (ML) has so rapidly integrated itself as a big player in the diagnostic fields for diabetes, several types of cancers (liver, thyroid, coronary, prostate, ovarian and skin) and stroke risk, which is the most pertinent to our project. The main overarching goal of the paper was to assess the difference in risk of bias (RoB) in ML and non-ML CVD predictors as such risks have not been accurately summarized since ML’s introduction. To accurately portray the RoB the researchers employed a PRISMA model for CVD studies and furthered those results using an analytical slope method.

The set of studies used to train the PRISMA model was parsed down to 117 studies after an initial 18,561 were identified, 15,000 were retained and then subjected to exclusion criteria. The main criteria across the steps were identified to exclude articles that were: non-relevant (14815 excluded), data was not given after screenings (62 excluded), or insufficient data for the training was present (excluded 6) which brought the total down to 117. The sites that the articles were drawn from included: “IEEE Xplore, PubMed, Google Scholar, and ScienceDirect” (Suri). The keywords used were: “‘CVD risk prediction using Artificial Intelligence’, ‘CVD risk assessment in AI framework’, ‘CVD risk prediction using carotid’, ‘CVD risk prediction using ML’, ‘CVD risk prediction’” (Suri). Once the model was adequately trained with these studies they were used to better analyze 117 articles of which 24 articles deemed to be pure ML and 14 articles deemed to be pure non-ML and the rest were a mixture of the two (60% to 40% pure ML to pure non-ML split). An important aspect to each of the pure ML studies was what ground truth design they used for their respective CVD calculator. The PRISMA model organized them into 5 distinct clusters: death (17%), coronary artery calcification (CAC) (24%), myocardial infarction (MI) along with stroke and angina (29%), diabetes along with hypertension (12%), and the event-equivalent gold standard (EEGS) (13%). There were a wide variety of different ML algorithms used in the study so only the most common ones will be mentioned which were: random forest (20%), decision tree (10%), deep learning (10%), K-nearest neighbor (5%), and neural network (5%) (Suri).

To better label the bias in these articles the researchers pre defined three “bias bins” that each study would be categorized in: low-bias (LB), moderate-bias (MB), and high-bias (HB). In order to properly place each study into the appropriate bin a cutoff value was calculated for each

bin by utilizing an already researched and credible concept of an AP ai bias score. This method requires attributes to be defined in each study and scored to take each article's individual score and divide by the total attributes. The researchers decided upon 46 different attributes for the bias score which was broken into nine categories for simplicity. These categories are “demographics, AI architecture design, optimization, performance parameters, scientific validation, statistical test, clinical validation, clinical evaluation, and survival and hazard analysis” (Suri). Using these attributes the cutoffs for the bis were calculated to be 1.83 for low to medium and 1.59 medium to high. To make it even more clear, the number of papers in each category for the pure ML articles was 5 for LB, 10 for MB, and 9 for HB. For the pure non-ML there were 3 LB, 4 MB, and 7 HB (Suri). Once the papers were categorized the researchers were able to determine the most pertinent characteristics of each paper that led to their placements. The LB studies were characterized with more sophisticated AI architecture and more polished classification metrics. The MB had less sophisticated AI architecture with a poorer method used for training for their data which resulted in worse cross-validation performance. The HB had the most bare bones AI architecture and had a clear lack of both clearly defined extraction processes and cross validation which in a general sense is indicative of a lack of clinical verification or validation.

Across both pure ML and pure non-ML the researchers found that there were some key demographic attributes that were the most commonly used, and it's no surprise that the data set we are using for our project has most of these attributes. In descending order of how many studies the attributes were found in: smoking (86%), hypertension (76.6%), non-image data (76.6%), BMI (70%), family history (63.4%), and ethnicity (38.4%) (Suri). It was surprising to see how relatively few articles had ethnicity in their data because it is often the starting point of discourse for bias in this type of research. Our own data does not have an ethnicity class either which has both advantages and disadvantages for introducing bias.

To better cement the results of the PRISMA model and the bin classification analysis, the researchers used the analytical slope method to further analyze the results. The analytical slope method is based on a proportion of scores (POS) that is obtained from the previously calculated AI attribute scores. There is a great deal of complicated math that goes into the score but the overall conclusion that the researchers came to was that bias in the ML papers were significantly less than the bias in the non-ML papers, a roughly 43% difference (Suri). This result might seem a bit surprising so the researchers made sure to emphasize the advantages to the approaches that they took. First was that since they used mathematical representations their results are more clear and straightforward. Second was that the slope method handled the differing proportions of ML vs non-ML represented in the data. Third was that they used a generalized technique instead of one they could have specifically designed to favor the outcome that they reached.

With all the relevant data analyzed and their results confirmed, the researchers were even better able to diagnose the causes in the bias for both ML and non-ML respectively. They found that for ML bias one of the most distinguishing factors between HB and LB was whether or not they included the smoking attribute. This result is a key reason why we have placed a great deal of focus in our project to cleaning and properly categorizing the smoking variable in the data we

have. Another distinguishing factor was whether or not the day they used came from verified and validated sources. The data set we have chosen reflects this factor as well since the data was collected from an official Electronic Health Record (EHR) overseen by McKinsey & Company which is a highly validated source. For the non-ML data the primary distinguishing factors for bias arose from a lack of clinical validation in the HB group.

Towards the end of the article, the researchers presented a list of primary and secondary recommendations to reduce RoB. The primary recommendations given to first choose the best outcome design for the model. As previously mentioned this was Death > CAC > Heart COnditions > Chronic Diseases > EGGS. The remaining primary suggestions were that if morphological images were used they should be based on phenotype risk factors, obtaining clinical evaluations were a must, minimize risk granularity, obtain cross modality validation, and have an adequate sample size with inter and intra observer variability. The secondary recommendations provided were centered around choosing the strongest model design that fits the desired output which included cross-validation, prediction, training, and graphical user interface (GUI). Beyond this, the secondary recommendations included choosing multiple data sets to reproduce results, obtaining multi-ethnic data sets, and opting for rigorous peer-review processes (Suri). Overall the paper provided key insights into a field that had gone unregulated for an extended period of time due to how fast it took over the predictor space. The results might have shown that the ML studies had less bias than the non-ML studies but the researchers urged the reader to understand that this difference did not totally quantify how biased the ML papers were on their own. This was in an active effort to remind readers that there is a great deal of work that still needs to be done to mitigate bias in both ML and non-ML models for which they provided valuable recommendations.