Anders Herlitz Health technology and algorithmic fairness

Many different kinds of new medical technology make use of predictive analytics in order to promote better health outcomes. Predictive analytics applies what I will call a predictive algorithm that is informed by big data on data about an individual to make predictions about the individual. When used in the healthcare sector, predictive analytics can help patients monitor their health-conditions and enable better self-medication, and it can improve the accuracy of medical diagnoses. However, it has recently been proven that unless the algorithm used to make predictions provides perfectly accurate predictions or the sought-after property (e.g. a health condition) is evenly distributed across groups, the predictive algorithm will violate at least one of three fairness axioms. Either (i) the algorithm correctly identifies the relevant property (e.g. a health condition) more often in one subgroup (e.g. men) than another (e.g. women); (ii) the algorithm produces more positive false findings of the property for one subgroup; or (iii) the algorithm produces more negative false findings for one subgroup. Otherwise put, predictive algorithms are in many contexts inherently discriminatory and exhibit the characteristics of wrongful discrimination that is non-intentional, but which results in different impacts on individuals based on what subgroup they belong to. This paper outlines and explores different responses to this problem: should we abandon predictive algorithms, compromise on equality in correct predictions across subgroups, or accept inequality between subgroups with respect to false findings? With reference to how the alternative costs of abandoning predictive algorithms are high and the results risk being even worse in terms of wrongful discrimination, that option is discarded as undesirable. The paper instead presents three ethical principles that should be universally applied to promote algorithmic fairness in relation to new health technology: (1) Transparency: those who promote the use of predictive algorithms should be aware of the unintended differences in impact and also be transparent to the affected community about these unintended differences; (2) Dominance: an algorithm that is better with respect to one of the three dimensions of fairness and worse with respect to none is better overall; and (3) Priority to the worse off: an algorithm that is relatively better for members of a worse-off subgroup is preferable to an algorithm that is relatively better for members of a better-off subgroup. Yet, the paper concludes that all-things-considered judgments about the fairness of predictive algorithms are context-dependent. They depend on the what is predicted, and what the negative consequences of false findings are.