Fairness Definitions for Digital Mental Health Applications

SEAMUS RYAN, School of Computer Science and Statistics, Trinity College Dublin, Ireland
GAVIN DOHERTY, School of Computer Science and Statistics, Trinity College Dublin, Ireland

Digital applications are becoming a standard part of the diagnosis, treatment, support of people dealing with mental health issues. While doing this they can potentially generate and use significant amounts of data about the user, their state, and their behaviour. Statistical models based on this data are likely to be used for a range of purposes, ranging from prediction of patient outcomes, to personalization of treatment. With such models, there comes a risk of introducing problems of inequality and bias. As unfair systems which lead to worse health outcomes for some groups of people have long existed in healthcare, such bias may emerge due to the historical data used to create models. Addressing fairness in the design of digital mental health tools requires a clear framework for defining and assessing fairness. In this paper we will categorise *fairness* definitions as applied in digital mental health. We move away from the existing models of grouping definitions based on likeness of measurement and towards a model of grouping based on likeness of design goal for a system. We propose three categories for fairness definitions that are reflective of healthcare concerns. We divide fairness definitions into the categories based on; 1) those that assume equal impact of error across users, 2) those that assume a homogeneous demographic, and 3) those that have a scope beyond the immediate decision. We then apply this categorisation to an existing case study.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**;

Additional Key Words and Phrases: Digital mental health interventions, Machine Learning, Fairness, Healthcare

ACM Reference Format:
Seamus Ryan and Gavin Doherty. 2021. Fairness Definitions for Digital Mental Health Applications. In *Proceedings of W35: Realizing AI in Healthcare: Challenges Appearing in the Wild (CHI '21).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Providing accessible and efficient mental heath support is an important goal within modern healthcare systems. Digital Health approaches are increasingly seen as an important strategy to augment traditional approaches to treatment. Approaches such as low intensity, internet based cognitive behavioral therapy (iCBT) have been demonstrated to be effective at improving health outcomes [33, 43]. Digital interventions can be used exclusively by clients, by clinicians, or by clients with clinician support. Existing examples of applications under these models include longitudinal psychological screening [14], supporting treatment [26], and medical history analysis [39].

While such applications can gather outcome data, they can also gather data regarding engagement with the intervention [30], or user behaviour [1, 5]. Some of this data is unique to digital delivery and would not be gathered during an exclusively face to face treatment. This data can be incorporated, where appropriate, in the guidance of medical treatment decisions with positive improvements to diagnosis and treatment [42]. This has led recent work in the health field to investigate the value of using Machine learning (ML), which can take data at this scale and produce statistical models to predict or guide healthcare decision making [38].

2 MACHINE LEARNING AND DIGITAL MENTAL HEALTH

The majority of ML analysis in healthcare has focused on anomaly detection in medical imagery (*e.g.* tumour detection in breast cancer imagery) [9, 23, 27]. Within mental health, researchers have begun

CHI '21, May 08–09, 2021, Online
2021. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

to explore the value of ML techniques [38]. Topics addressed have included diagnostic guidance in clinical depression [32], the prediction of the onset of mental illnesses [37], identification of patients off track for treatment progression [13], momentary interventions [4], and the prediction of bouts of negative emotions among otherwise healthy users [19].

It is important to be aware that ML is a broad term that includes a range of different algorithmic approaches to the analysis of data. This data can can be categorical (*e.g.* sex, scores of psychometric scales, time between or since medical visits, family history, etc.) or in a form that is continuous (*e.g.* application usage behaviour, text mood logs, movement patterns, etc.). It is possible to incorporate data into a ML model that would not be included in an analog clinical decision making process due to it's scale or the time required to analyse it manually. The inclusion of these variables does not mean that they will be given weight in the process, but does allow for the identification of potentially overlooked, medically relevant trends.

Randomised Control Trials are seen as being the most appropriate benchmark to use when measuring the efficacy of a health treatment [7]. However, few ML models have seen such validation with real patients, and limited examples exist of ML adoption within medical environments [35]. Most ML studies in healthcare tend to be experimental, retrospective, and isolated from the larger workflow that exists in the healthcare system.

Machine learning models typically take historical data as ground truth. Predictions are then made based on models created using this data. If this historical data contains examples of discriminatory decisions, then the model created will be discriminatory. An example of this is the historical under-diagnosis of mental health issues in children with a minority background [24]. Also, if the data does not accurately represent the full range of patient demographics, then the baseline for normative data will be skewed towards the attributes of the over-represented group (*e.g.* lack of gender diversity in biomedical research [44]). This historical bias, and biases that are created by lack of representation are two ways that data can lead to unfair outcomes [28]. As the growing potential for ML adoption becomes clearer, analysing how it can be deployed into the real world safely and fairly has become a significant area of research [28].

2.1 Fairness in Healthcare

As unfair systems in healthcare can lead to significantly poorer patient outcomes [2, 3, 18, 36, 44], the need for methods for ensuring fairness in healthcare is clear. However this domain raises unique and complex challenges; Healthcare has documented issues of inequality of treatment [2, 18, 36] as well as a history of intentional trade-offs to gain efficiency at the cost of *fairness* [34]. Given this background, as well as the high-stakes nature of diagnosis and treatment, there needs to be a clear understanding of what is meant by *fairness* in order to identify and minimise biases in healthcare software.

Introducing measures of fairness into Healthcare is complex as the same treatment can contribute towards different outcomes depending on the patient. Some patients may follow a course of treatment where the outcome for them is improved symptom management and for other patients the outcome of the same treatment may be full remission. This can be thought about as equality vs equity. A system that is focused on ensuring the patients are treated the same is concerned with equality (**e.g.** ensuring equal accuracy of diagnosis). This is distinct from a system that focuses on long term patient outcomes, referred to as equity (*e.g.* varying the treatment plan to achieve equal outcomes for different groups of people). In any practical patient experience with a sufficiently complex healthcare process, both equality and equity will be important at different times. Using a fairness definition that looks at the equality of the immediate treatment where the long term equity of the treatment process is more important will lead to seemingly equivalent patient care but unequal long term outcomes.

There should also be an awareness of the limitations of using metric driven approaches in medicine. These measures are derived from the comparison of correct vs incorrect outcomes. For example, assessments based on comparing the sum of Type 1 errors, referred to in the Machine Learning literature as False Positives, presupposes that the overall impact of Type 1 errors is equitable between groups [6]. In healthcare this is not true as misdiagnosis or incorrect treatment guidance can have a disproportionate impact on marginalised groups [24]. Because of these concerns, definitions of Fairness that use equality or conditional equality of false positives may not be appropriate in areas with divergent long term outcomes for similar decisions.

A number of frameworks for fairness have been proposed that aim to bridge the gap between mathematical performance and real-world impact; These include the *AI Fairness Checklist* [25] which guides an organisation through a fair co-design process, and Mitchell et al [29], who elaborate on the decisions inherent to the design of tailored systems. Both of these depend on proactive planning and management in the analysis of fairness concerns.

2.2 Fairness in Digital Mental Health

In digital mental health, the generation of large data-sets and possibilities to directly deploy ML tools make it an area with a lot of potential for personalized delivery of treatment. However the creation of *"intervention-generated inequalities"* [40] is possible in all areas of healthcare, and as personalized DMH tools scale, so will these risks.

Fairness considerations exist in all aspects of mental health support, however one of the benefits of the digital mental health environment is that there is potentially less scarcity of resources to allocate. In other aspects of support, limited availability of clinician time or places on programs may impose strong constraints. While no digital application can scale indefinitely without additional engineering and organizational processes, the allocation of digital support to one person will typically have a low impact on the availability of this resource to others. A prerequisite for analysis of fairness with respect to demographics is having the demographic information. Traditionally this has only been part of the data-set used in digital health applications when directly relevant to the goal of the intervention due to the sensitive nature of the data. Thus a potential design tension exists between data minimisation strategies to protect privacy, and ensuring fairness.

3 DEFINITION ANALYSIS OF FAIRNESS FOR DIGITAL MENTAL HEALTH

In the 2018 systematic review of fairness definitions *"Fairness definitions explained"*[41], Verma and Rubin detail 20 different definitions (Appendix Table 1) broken into 3 categories. These categories comprise Statistical measures, Similarity-based measures, and Definitions based on causal reasoning. Designers working on applications that leverage ML face the challenge that definitions are grouped based on the mathematical origin or interpretability, not by what they actually mean to end users. Definitions of *fairness* can be contradictory and, as the field continues to expand and definitions are refined, it will be part of the engineering and design process to decide what it will mean for particular software applications.

In order to help designers to understand the meaningful impact a fairness definition will have on their users we propose categorising definitions based on the requirements and expectations of the users. By focusing on design requirements for definitions, such an approach could provide some meaningful distinctions of relevance to the design of digital health applications.

Error Aware. Error aware definitions are those that compare error rates between demographic groups. For the design of some medical applications, errors are not equal. Failing to detect a medical issue (a false negative (FN)) is not the same as incorrectly suggesting there is one when there is not (a false positive (FP)); In the same way, failing to detect depression in a vulnerable patient is not

the same as failing to detect depression in a patient with a strong support network; Both are false negatives but they are not equal to each other. In some areas of healthcare we may aim for equal error rates, such as the ability to parse a physical medical record and a digital medical record for relevant information. Error aware definitions are needed when the designers of a process feel that the focus should be on achieving similar error rates for different groups (rather than on the impact of errors), and that these errors should be minimised.

Demographic aware. Demographic aware definitions are those that examine the performance of a process for different demographic groups. For example, those that compare the sum of positive results between demographic groups, or expect the demographic proportions of the cohort receiving positive results to reflect a population. In some medical decisions we do not expect equal levels of occurrence between different groups; Among refugees we might expect to see higher occurrence levels of PTSD compared to the broader population, for example. We may also want to take demographics into account to ensure equitable access to resources; for example, ensuring that people within different regions have similar access to primary healthcare facilities. Demographic aware definitions are needed when the designers of a process feel that all demographic groups need to be represented equally or proportionally in an outcome or decision.

Impact Aware. Impact Aware definitions are those that incorporate the long term impact a decision can have on a person. In medicine the same treatment can lead to a different long term impact for different patients or groups of patients. This is common in medication prescription where different quantities or types of medication may be suggested based on age or body type for the same issue. In these cases the treatment is different but the long term impact is the same. In cognitive behaviour therapy patients may start the same program but where one might achieve recovery for another it may serve to avoid deterioration. In these cases the treatment is the same but the long term impact is different. Impact aware definitions are needed when the designers of a process expect patients to be given different decisions but build towards the same long term impact.

Taking existing definitions of fairness [41] it is possible to identify which meet these criteria (Appendix Table 1).

4 SCENARIO ANALYSIS OF FAIRNESS FOR DIGITAL MENTAL HEALTH

We apply these categories to an existing scenario of ML in mHealth by way of illustration, and with the goal of exploring the issues involved.

4.1 Scenario - ML analysis of Support communications to clients doing CBT modules

4.1.1 Scenario Description. An ML model was used to identify the types of clinical support messages that had the greatest impact on clients using an online mental health intervention [10]. Based on their messages, supporters were clustered into groups identified as having a high, medium, and low impact on a set of defined patient healthcare outcomes. The text used in messages from supporters was then analysed for linguistic markers. This workflow is mapped in figure 1.

4.1.2 Appropriate Fairness models. As this work is looking at improvement of an existing process rather then the tailoring of behaviour, there are no between group comparisons inherent to the research. Instead we can define fairness from the point of view of the individuals being acted on, in figure 1 these are identified as *s1.1*, the supporter, and *s1.2*, the client.

The inputs to the ML model are clinical outcomes at the client level (outcome change over time) and at the message level (outcome changes following messages). The model then groups supporters into clusters, based on the impact they have had on client outcomes. For the supporter (s1.1), the decision to group a them into one cluster does not have any long term impact. As such an *Impact*

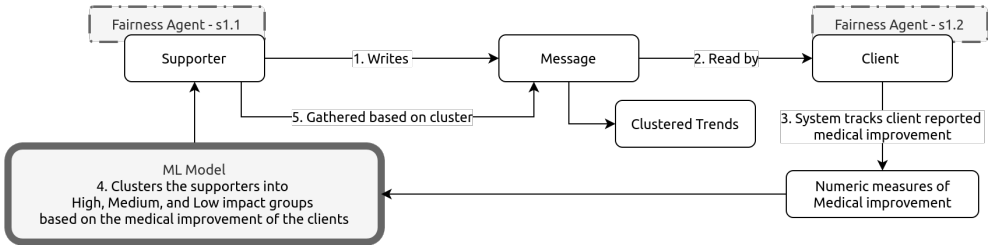

Fig. 1. Simplified flow of data inputs and parameters

aware definition of fairness is not required for the analysis of the effect on s1.1. There is no expected demographic divergence in the clustering, as we would not expect any one demographic group to have a higher or lower impact. As such a definition of fairness does not need to be *Demographic aware*.

It is possible that the measurement of change of s1.2 well-being may have lower level of accuracy for members of demographic groups due to cultural norms in the treatment or reporting of mental health concerns. This would lead to certain demographics being over or underrepresented in the high impact cluster leading to the later psycho-linguistic analysis being weighted closer to the normative styles of communication of a given demographic group. Assuming that clients are randomly assigned to supporters then regression to the mean will offset this concern, however it is a possible bias in the data. As such we need to ensure an accuracy focused definition of error awareness is met.

The second agent in the model is the client receiving messages (*s1.2* in figure 1). As the model is not clustering the clients, there is less chance for misrepresentation or over-representation of a cohort. However, this assumes the gathered clinical data do not contain any biases. If, for example, one of the scales used in this analysis has lower accuracy for a given demographic, then historical bias may be present in the input data-set, and could have an effect on the model outputs. For this reason an *Error aware* definition is appropriate.

4.1.3 Discussion. As this machine learning model is focused on improvement of an existing treatment process, and does not involve the tailoring or distinguishing of user groups, there are limited Demographic or Impact fairness issues concerning its use and deployment. Regardless of the presence of demographic data from supporters, the possibility of data bias still exists due to proxy variables. Based on this the requirement for an Error Aware definition of fairness is clear. As data collected directly from supporters is not used in the clustering of the model then certain conditions for fairness, (e.g. *Fairness through unawareness* (ID 15 Appendix - Table 1), are already met. The scenario detailed here would have had different fairness concerns had the machine learning model clustered the messages instead of supporters. This approach would have had significantly greater fairness concerns as demographic and outcome aware definitions would be needed to design for equitable treatment of different demographic groups.

5 CONCLUSION

Researching and proposing fairness models for the design of digital mental health now will support the design of applications that work for everyone. In this short paper we have considered the issue of fairness within the context of digital mental health, and some of the challenges specific to this domain. We have proposed a categorisation of fairness models which is oriented towards design, and explored this through a case study.

6 ACKNOWLEDGMENTS

This work has been supported in part by Science Foundation Ireland under Grant number 18/CRT/6222.

REFERENCES

[1] Saeed Abdullah, Mark Matthews, Ellen Frank, Gavin Doherty, Geri Gay, and Tanzeem Choudhury. 2016. Automatic detection of social rhythms in bipolar disorder. *Journal of the American Medical Informatics Association* 23, 3 (5 2016), 538–543. https://doi.org/10.1093/jamia/ocv200

[2] Corey M. Abramson, Manata Hashemi, and Martín Sánchez-Jankowski. 2015. Perceived discrimination in U.S. healthcare: Charting the effects of key social characteristics within and across racial groups. , 615–621 pages. https://doi.org/10.1016/j.pmedr.2015.07.006

[3] K. Beery Annaliese and Zucker Irving. 2011. Sex Bias in Neuroscience and Biomedical Research. *Neurosci Biobehav Rev* 23, 1 (2011), 1–7. https://doi.org/10.1038/jid.2014.371

[4] Andreas Balaskas, Stephen M. Schueller, Anna L. Cox, and Gavin Doherty. 2021. Ecological momentary interventions for mental health: A scoping review. *PLOS ONE* 16, 3 (03 2021), 1–23. https://doi.org/10.1371/journal.pone.0248152

[5] Jakob E. Bardram, Mads Frost, Károly Szántó, and Gabriela Marcu. 2012. The MONARCA self-assessment system. In *Proceedings of the 2nd ACM SIGHIT symposium on International health informatics - IHI '12*. ACM Press, New York, New York, USA, 21. https://doi.org/10.1145/2110363.2110370

[6] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* 50, 1 (2 2021), 3–44. https://doi.org/10.1177/0049124118782533

[7] Iain Chalmers, Murray Enkin, and Marc J. N. C. Keirse. 1993. Preparing and Updating Systematic Reviews of Randomized Controlled Trials of Health Care. *The Milbank Quarterly* 71, 3 (1993), 411. https://doi.org/10.2307/3350409

[8] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness Under Unawareness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 339–348. https://doi.org/10.1145/3287560.3287594

[9] Jonathan H Chen, Steven M Asch, and Palo Alto. 2018. Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. *N Engl J Med* 376, 26 (2018), 2507–2509. https://doi.org/10.1056/NEJMp1702071.Machine

[10] Prerna Chikersal, Danielle Belgrave, Gavin Doherty, Angel Enrique, Jorge E Palacios, Derek Richards, and Anja Thieme. 2020. Understanding Client Support Strategies to Improve Clinical Outcomes in an Online Mental Health Intervention. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376341

[11] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. https://doi.org/10.1089/big.2016.0047

[12] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* Part F1296 (2017), 797–806. https://doi.org/10.1145/3097983.3098095

[13] Jaime Delgadillo, Kim de Jong, Mike Lucock, Wolfgang Lutz, Julian Rubel, Simon Gilbody, Shehzad Ali, Elisa Aguirre, Mark Appleton, Jacqueline Nevin, Harry O'Hayon, Ushma Patel, Andrew Sainty, Peter Spencer, and Dean McMillan. 2018. Feedback-informed treatment versus usual psychological treatment for depression and anxiety: a multisite, open-label, cluster randomised controlled trial. *The Lancet Psychiatry* 5, 7 (7 2018), 564–572. https://doi.org/10.1016/S2215-0366(18)30162-7

[14] Kevin Doherty, Andreas Balaskas, and Gavin Doherty. 2020. The Design of Ecological Momentary Assessment Technologies. *Interacting with Computers* 00, 0 (8 2020), 1–12. https://doi.org/10.1093/iwcomp/iwaa019

[15] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*. ACM Press, New York, New York, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[16] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: Testing software for discrimination. *Proceedings of the ACM SIGSOFT Symposium on the Foundations of Software Engineering* Part F1301 (2017), 498–510. https://doi.org/10.1145/3106237.3106277 arXiv:1709.03221

[17] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems* (10 2016), 3323–3331. http://arxiv.org/abs/1610.02413

[18] Leslie R.M. Hausmann, Nancy R. Kressin, Barbara H. Hanusa, and Said A. Ibrahim. 2010. Perceived racial discrimination in health care and its association with patients' healthcare experiences: Does the measure matter? *Ethnicity and Disease* 20, 1 (2010), 40–47.

[19] Galen Chin-Lun Hung, Pei-Ching Yang, Chia-Chi Chang, Jung-Hsien Chiang, and Ying-Yeh Chen. 2016. Predicting Negative Emotions Based on Mobile Phone Usage Patterns: An Exploratory Study. *JMIR Research Protocols* 5, 3 (2016),

e160. https://doi.org/10.2196/resprot.5551

[20] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems* 2017-Decem, Nips (2017), 657–667.

[21] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. *Leibniz International Proceedings in Informatics, LIPIcs* 67 (2017), 1–23. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[22] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in Neural Information Processing Systems* 2017-Decem, Nips (2017), 4067–4077.

[23] Steven Lemm, Benjamin Blankertz, Thorsten Dickhaus, and Klaus Robert Müller. 2011. Introduction to machine learning for brain imaging. *NeuroImage* 56, 2 (2011), 387–399. https://doi.org/10.1016/j.neuroimage.2010.11.004

[24] June Liang, Brittany E. Matheson, and Jennifer M. Douglas. 2016. Mental Health Diagnostic Considerations in Racial/Ethnic Minority Youth. *Journal of Child and Family Studies* 25, 6 (2016), 1926–1940. https://doi.org/10.1007/s10826-015-0351-z

[25] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376445

[26] Mark Matthews and Gavin Doherty. 2011. In the Mood: Engaging Teenagers in Psychotherapy Using Mobile Phones. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*. ACM Press, New York, New York, USA, 2947. https://doi.org/10.1145/1978942.1979379

[27] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C. Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona J. Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher J. Kelly, Dominic King, Joseph R. Ledsam, David Melnick, Hormuz Mostofi, Lily Peng, Joshua Jay Reicher, Bernardino Romera-Paredes, Richard Sidebottom, Mustafa Suleyman, Daniel Tse, Kenneth C. Young, Jeffrey De Fauw, and Shravya Shetty. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577, 7788 (2020), 89–94. https://doi.org/10.1038/s41586-019-1799-6

[28] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. *ArXiv* abs/1908.0 (8 2019). http://arxiv.org/abs/1908.09635

[29] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8, 1 (3 2021), 042720–125902. https://doi.org/10.1146/annurev-statistics-042720-125902

[30] Cecily Morrison and Gavin Doherty. 2014. Analyzing engagement in a web-based intervention platform through visualizing log-data. *Journal of Medical Internet Research* 16, 11 (2014), 1–16. https://doi.org/10.2196/jmir.3575

[31] Razieh Nabi and Ilya Shpitser. 2018. Fair Inference On Outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32. 1931–1940. http://arxiv.org/abs/1705.10378

[32] Anastasia Pampouchidou, Panagiotis G. Simos, Kostas Marias, Fabrice Meriaudeau, Fan Yang, Matthew Pediaditis, and Manolis Tsiknakis. 2017. Automatic Assessment of Depression Based on Visual Cues: A Systematic Review. *IEEE Transactions on Affective Computing* 10, 4 (2017), 445–470. https://doi.org/10.1109/TAFFC.2017.2724035

[33] D. Richards, L. Timulak, E. O'Brien, C. Hayes, N. Vigano, J. Sharry, and G. Doherty. 2015. A randomized controlled trial of an internet-delivered treatment: Its potential as a low-intensity community intervention for adults with symptoms of depression. , 20–31 pages. https://doi.org/10.1016/j.brat.2015.10.005

[34] Jeff Richardson and Michael Schlander. 2019. Health technology assessment (HTA) and economic evaluation: efficiency or fairness first. *Journal of Market Access & Health Policy* 7, 1 (2019), 1557981. https://doi.org/10.1080/20016689.2018.1557981

[35] Adrian B. R. Shatte, Delyse M. Hutchinson, and Samantha J. Teague. 2019. Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine* 49, 09 (7 2019), 1426–1448. https://doi.org/10.1017/S0033291719000151

[36] Anna Skosireva, Patricia O'Campo, Suzanne Zerger, Catharine Chambers, Susan Gapka, and Vicky Stergiopoulos. 2014. Different faces of discrimination: Perceived discrimination among homeless adults with mental illness in healthcare settings. *BMC Health Services Research* 14, 1 (2014), 1–11. https://doi.org/10.1186/1472-6963-14-376

[37] M. Srividya, S. Mohanavalli, and N. Bhalaji. 2018. Behavioral Modeling for Mental Health using Machine Learning Algorithms. *Journal of Medical Systems* 42, 5 (5 2018), 88. https://doi.org/10.1007/s10916-018-0934-5

[38] Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine Learning in Mental Health. *ACM Transactions on Computer-Human Interaction* 27, 5 (10 2020), 1–53. https://doi.org/10.1145/3398069

[39] Truyen Tran, Wei Luo, Dinh Phung, Richard Harvey, Michael Berk, Richard L. Kennedy, and Svetha Venkatesh. 2014. Risk stratification using data from electronic medical records better predicts suicide risks than clinician assessments. *BMC Psychiatry* 14, 1 (2014), 1–9. https://doi.org/10.1186/1471-244X-14-76

[40] Tiffany C. Veinot, Hannah Mitchell, and Jessica S. Ancker. 2018. Good intentions are not enough: how informatics interventions can worsen inequality. *Journal of the American Medical Informatics Association* 25, 8 (8 2018), 1080–1088. https://doi.org/10.1093/jamia/ocy052

[41] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*. ACM, New York, NY, USA, 1–7. https://doi.org/10.1145/3194770.3194776

[42] David S. Watson, Jenny Krutzinna, Ian N. Bruce, Christopher E.M. Griffiths, Iain B. McInnes, Michael R. Barnes, and Luciano Floridi. 2019. Clinical applications of machine learning algorithms: Beyond the black box. *BMJ (Online)* 364, March (2019), 10–13. https://doi.org/10.1136/bmj.l886

[43] Michael J. Wells, Jesse J. Owen, Laura W. McCray, Laura B. Bishop, Tracy D. Eells, Gregory K. Brown, Derek Richards, Michael E. Thase, and Jesse H. Wright. 2018. Computer-Assisted Cognitive-Behavior Therapy for Depression in Primary Care. *The Primary Care Companion For CNS Disorders* 20, 2 (3 2018). https://doi.org/10.4088/PCC.17r02196

[44] Nicole C Woitowich, Annaliese Beery, and Teresa Woodruff. 2020. A 10-year follow-up study of sex inclusion in the biological sciences. *eLife* 9 (6 2020), 1–8. https://doi.org/10.7554/eLife.56344

A DEFINITIONS FROM "FAIRNESS DEFINITIONS EXPLAINED"

Fairness Name	ID	Explanation	Error Aware	Demo Aware	Impact Aware
Statistical Parity [15]	1	Statistical Parity is an algorithmic definition and seeking likelihood of a "positive" outcome is the same for those with Protected Attributes as unprotected.	**N**	**Y**	**Y**
Conditional Statistical Parity [12]	2	Conditional Statistical Parity is an algorithmic definition that extends the statistical parity metric to allow fluctuations in likelihood if there is "Legitimate factors"	**N**	**!**	**Y**
Equal Opportunity [11]	3	Predictive Parity is an algorithmic definition that looks at actual results and defines fair as equal Positive Predictive Value (PPV) between all Groups. Positive Predictive Value is the likelihood of a positive prediction to be accurate (True positives / True positives + False positives).	**Y**	**N**	**N**
False Positive Error Rate [11]	4	False Positive Error Rate is an algorithmic definition that looks at actual results and defines fair as equal False Positive Rate (FPR) between all Groups. False Positive Rate is the likelihood of a negative prediction to be given, incorrectly, as positive (False positives/False positives + True Negative)	**Y**	**N**	**N**
False Negative Error Rate [11]	5	False Negative Error Rate is an algorithmic definition that looks at actual results and defines fair as equal False Positive Rate (FNR) between all Groups. False Negative Rate is the likelihood of a negative prediction to be given, incorrectly, as positive (False Negative /False Negative + True Positive)	**Y**	**N**	**N**
Equalised Odds [17]	6	Equalised Odds is an algorithmic definition that combines False Positive Error Rate and Equal Opportunity definitions that looks to ensure that the PPV and FPR rate are equal across demographic groups.	**Y**	**N**	**N**
Conditional Use Accuracy Equality [6]	7	Conditional Use Accuracy Equality is an algorithmic definition, similar to Equalised Odd, it focuses on equal levels of positive predictive value (TP/ TP+ FP) and Negative Predictive Value (TN / TN+FN) across demographic groups.	**Y**	**N**	**N**
Overall Accuracy Equality [6]	8	Overall Accuracy Equality is an algorithmic definition that focuses on accurate of the true positive and true negative results across demographic groups (TP = TP, TN = TN)	**Y**	**N**	**N**

Fairness Name	ID	Explanation	Error Aware	Demo Aware	Impact Aware
Treatment Equality [6]	9	Treatment Equality is an algorithmic definition that looks at errors as the standard for parity. If FP and FN are equal between these groups then fairness is met	**Y**	**N**	**N**
Test-fairness or calibration [17]	10	Test-fairness is both an algorithmic and outcome based definition that looks at both the positive predictive value as being equal across all groups and that the fraction of correct positive prediction being the same in all groups.	**Y**	**Y**	**N**
Well-Calibration [21]	11	Well-Calibration is an algorithmic definition of fairness that extends Equal Opportunity (Predictive Parity) to include the likelihood of receiving a positive score should be identical for all.	**Y**	**Y**	**N**
Balance for Positive Class [21]	12	Balance for Positive Class is an algorithmic definition that looks for an equal TP between groups.	**N**	**Y**	**N**
Balance for Negative Class [21]	13	Balance for Negative Class is an algorithmic definition that looks for an equal TN between groups.	**N**	**Y**	**N**
Causal discrimination [16]	14	Causal discrimination is a similarity based definition that looks at specific causal attributes. The decision should be that same for any two subject with the same attributes.	**N**	**N**	**N**
Fairness through unawareness [8]	15	Fairness through unawareness is a conceptual frame for decision making models the act where any Protected Variables are intentionally excluding from the model. This could be done at the pre processing stage.	**N**	**N**	**N**
Fairness through awareness [15]	16	Fairness through awareness is a conceptual frame for decision making models which includes the context of the wider social and historical ecosystem to ensure historic patterns of bias are not reflected. Protected variables cannot be excluded. It extends statistical parity to include minimising distance between subsets within demographics.	**N**	**Y**	**N**
Counterfactual Fairness [22]	17	Counterfactual Fairness is an algorithmic definition that focuses on taking a result from a model and considering it fair if the result is the same when the Protected Variables are changed.	**N**	**Y**	**N**
No unresolved discrimination [20]	18	No unresolved discrimination is a graph based fairness definition that looks to ensure that there are "resolving nodes" between a Protected Node and the decision node. A resolving node is one that is influenced by a Protected Variable but not in a way that moves away from the otium outcome.	**N**	**N**	**!**
No proxy discrimination [20]	19	No proxy discrimination is a conceptual definition and acts as an augmentation to other forms of fairness that is true if there is no path between a proxy variable node and the decision node when represented on a causal diagram.	**N**	**N**	**!**
Fair Inference [31]	20	Fair Inference is a graph based fairness definition and is similar to "No unresolved discrimination". It focuses on groups of nodes and vertices that it refers to as "legitimate" or "Illegitimate". Fair inference is satisfied if no Illegitimate paths are used to reaching the decision.	**N**	**N**	**!**

Table 1. Parent Table of definition Summary