ADVANCECRT

Health Technology Design

Centre for Research Training

Science Foundation Ireland sfi For what's next

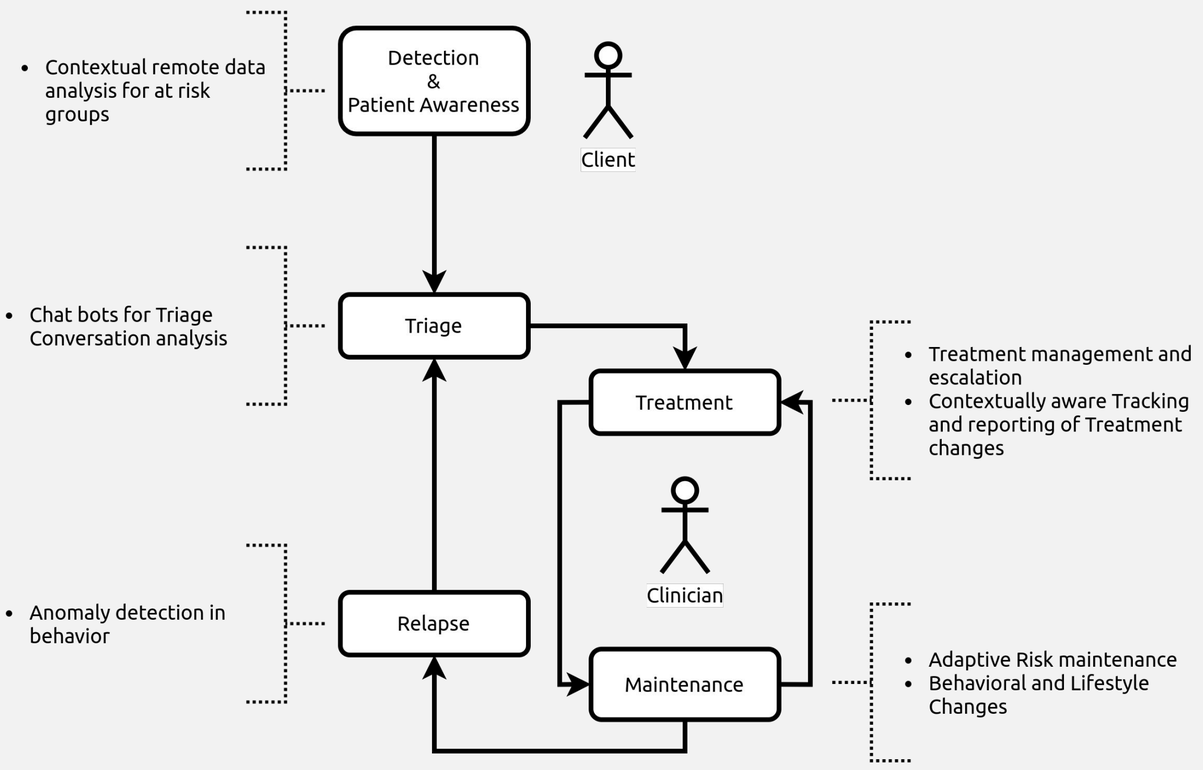
Designing Fair Machine Learning and Artificial Intelligence in Digital Mental Health

- Seamus Ryan (ryans58@tcd.ie) , Supervised by Dr Gavin Doherty

Digital Mental Health is in a strong position to take advantage of the benefits of Machine learning and Artificial intelligence. Large quantities of longitudinal data are generated during digital health interventions, and as the field to beings to adopt the tools of advanced statistical analysis on this data there is many areas in which client support can benefit.

Healthcare also contains a logical succession of potential implementations that lead from low risk areas, such as the timing of meditation reminders, to areas requiring high or perfect accuracy such as medication reminders. The practices, models, and safeguarded developed for one logically support the research of the next.

Mental health services are more in demand than ever and the data used by people in there day to day life, generated by IOT devices, could become an important part of how a modern health care system supports patients. These digital services should be met with clear and transparent analysis of their safety and efficiency. Key to this is analysis of **Fairness** in their decisions.



Fairness Definitions			
Statistical Parity	Equalised Odds	Well-Calibration	Fairness through awareness
Conditional Statistical Parity	Fairness through unawareness	Balance for Positive Class	Counterfactual Fairness
Predictive Parity	Overall Accuracy Equality	Balance for Negative Class	No unresolved discrimination
Test-fairness or calibration	Treatment Equality	Causal discrimination	No proxy discrimination
False Negative Error Rate Balance	False Positive Error Rate Balance	Conditional Use Accuracy Equality	Fair Inference

Verma, S., & Rubin, J. (2018). **Fairness definitions explained**. Proceedings - International Conference on Software Engineering, 1–7. https://doi.org/10.1145/3194770.3194776

Q . With ~25+ definitions of Fairness being incorporated into ethical standards how does a researcher build AI/ML into Medical Applications?

A . Very Carefully

Fairness is the term used for practices of ensuring that no person or group of people is treated worse than any other. This, along with Transparency and Accountability, is the underpinning for how most researchers approach safe and ethical Machine Learning.

Many common definitions of **Fairness** focus on the statistical parity between different demographic groups. These parity metrics use identified variables on which there is a risk of discrimination (examples being gender, ethnicity, sexual orientation). These definitions are appropriate in some areas of digital mental health but can be absurd to consider in others. Imagine a Healthcare system focused on postpartum mental health that looks for parity between genders.

Fairness needs to be treated as more than a statistical or Machine Learning problem but as a sociological and designed issue. There are tools in Human-Computer Interaction field the can potentially be used in the resolution of issues like this the, however work needs to done to find their role and the language we need when talking about them.

Designing for Fair in practice

Discussion on Fairness have recently seen a trend in moving from abstracts conceptual definitions towards specific domain data applications. In the analysis of Digital Mental Health it is beneficial in conversation to talk using examples.

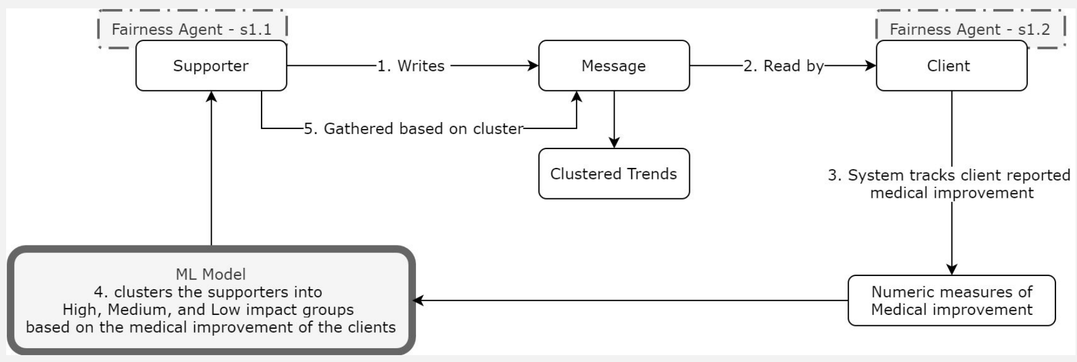


Fig 1 - Adapted from Chikersal, P., Belgrave, D., Doherty, G., Enrique, A., Palacios, J. E., Richards, D., & Thieme, A. (2020). **Understanding Client Support Strategies to Improve Clinical Outcomes in an Online Mental Health Intervention.** Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, (February), 1–16. https://doi.org/10.1145/3313831.3376341

Taking the example from the paper *Understanding Client Support Strategies to Improve Clinical Outcomes in an Online Mental Health Intervention* (Chikersal et al) detailed above, we can see fairness requiring analysis from two points of view, that of the supporter, sending the messages to the clients, and the clients themselves.

Both of these cohorts have concerns in a developed Machine Learning model. Were we designing an website we would take both concerns into accounts, this can not and should not change for ML.

This is my current work, finding ways to include Fairness in the design process.

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Host Institution
UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

MTU
Ollscoil Teicneolaíochta na Mumhan
Munster Technological University

Maynooth University
National University of Ireland Maynooth

TU DUBLIN
Ollscoil Teicneolaíochta Bhaile Átha Cliath
Technological University Dublin