# Developing a Fair AI-based Healthcare Framework with Feedback Loop

Vithya Yogarajan[1,*], Gillian Dobbie[1], Sharon Leitch[2] and David Reith[3]

[1]School of Computer Science, The University of Auckland, New Zealand
[2]General Practice and Rural Health, Otago Medical School, University of Otago, Dunedin, New Zealand
[3]Office of the Dean, Otago Medical School, University of Otago, Dunedin, New Zealand

Abstract
Artificial intelligence (AI) driven technological advances have increased the concerns over the bias problem, especially in high stake applications such as healthcare. This research proposes an AI-based healthcare framework with a feedback loop to ensure a fair real-world practical solution. We argue the need to consider continuous quality improvements of such technologies by including feedback at each development and deployment stage. Furthermore, medical experts, AI experts and policymakers must work together to ensure fairness. We use simple New Zealand-based patient data as a case study. We provide early-stage experimental results using machine learning algorithms where fairness and mitigation are also considered in addition to accuracy measures.

**Keywords**
Healthcare framework, Artificial intelligence, Continuous quality improvements, Bias, Fairness

## 1. Introduction

In a technological era driven by artificial intelligence (AI), the benefits of incorporating such advances in real-world high stake applications, including healthcare, are becoming the norm [1, 2]. Equally, there is an increase in concerns and awareness of the issues related to the bias problems in AI models [3, 4, 5]. In general, fairness is seen as being impartial and fair. AI models can be biased and can make "unfair" decisions, where such decisions are skewed toward a particular group of people [6, 7]. While there is an increase in urgency towards handling the bias problem and developing fair AI models, especially in clinical settings, such research is limited by the need to consider practical deployment. Furthermore, predominant research is focused on the Black and White racial issues in the US population [8, 7].

Recent studies have emphasised a need for continual monitoring and update to ensure the long-term reliability and effectiveness of AI-based clinical algorithms [9, 10]. Continuous quality improvement (CQI) [11] is a common phenomenon in healthcare, where incremental and progressive improvements are considered at various stages of operations to ensure patient care and safety. As such, using AI in clinical settings to aid clinical decisions and risk predictions requires CQI.

This paper proposes developing an AI-based healthcare framework incorporating mechanisms to ensure fairness and consider CQI. We emphasise that at each stage of the process –data collection, algorithm development, evaluation and implementation– there is a need to include a feedback loop where AI experts, medical professionals and policymakers are involved. We also provide a simple real clinical example in the New Zealand (NZ) setting to demonstrate the stages of the framework. The research presented in this paper is at the development stage, and as such, we acknowledge there are limitations and scope for improvements.

## 2. Framework

Figure 1 provides an overview of the proposed framework. We have split the process into four parts, (i) data, (ii) problem definition, (iii) algorithm selection and (iv) feedback, and also list the vital stages of each part. We use dash lines to show connections between the parts where feedback is beneficial and may include several cycles, and arrows indicate the flow's direction. Figure 1 also provides experts' level of involvement at each stage, and decisions that require predominant input from policymakers are provided in 'red'. In this framework, we view patients as stakeholders, and as such, their engagements can improve research appropriateness, acceptability, feasibility, delivery, and dissemination [12]. In New Zealand, the indigenous population Māori and other minority groups are important stakeholders, so it is crucial to handle their data with care [13, 14, 15] and incorporate feedback from the minority groups.

### 2.1. Fairness measures

Standard group fairness measures include demographic parity, disparate impact, equalized odds, and equalized

*The 6th international workshop on Knowledge Discovery in Healthcare Data (KDH), August 20, 2023, MACAO, China*
vithya.yogarajan@auckland.ac.nz (V. Yogarajan)
https://profiles.auckland.ac.nz/vithya-yogarajan (V. Yogarajan)
0000-0002-6054-9543 (V. Yogarajan); 0000-0001-7245-0367 (G. Dobbie); 0000-0001-9939-8773 (S. Leitch)

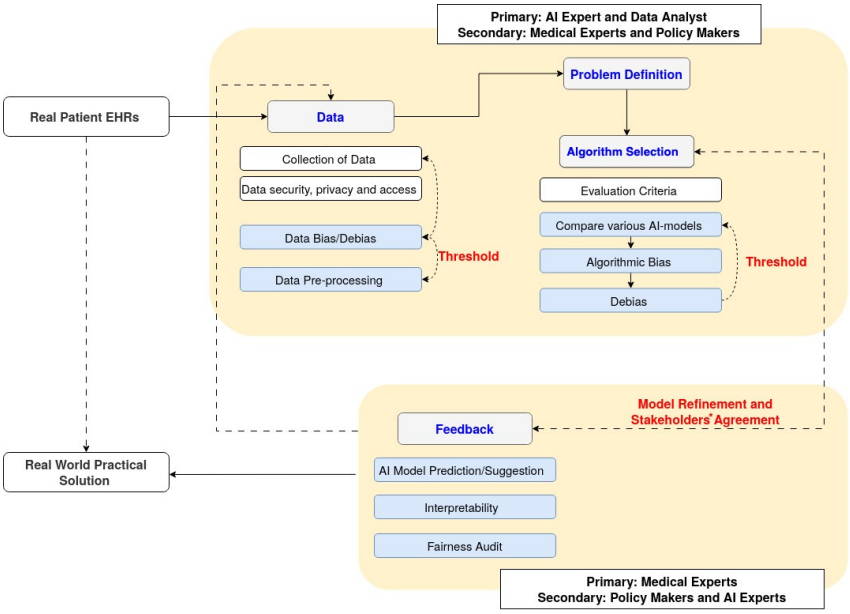



**Figure 1:** Flowchart of a fair AI-based healthcare framework with feedback loop at each stage. The dash lines show connections between parts where feedback is beneficial, and arrows indicate the flow's direction. *Patients are viewed as stakeholders.

opportunity, where fairness concerns a group rather than an individual [16]. In this paper, we use disparate impact (DI) [17] as a quantitative measure of fairness. DI measures the ratio of rates at which the outcomes occur for one group of patients over the rest [17, 16, 18]. According to US legislation, the $DI > \tau$ threshold was set at $\tau = 0.8$. While in practice, the acceptable range for the DI ratio is generally between 0.8 and 1.25. It is vital to point out that setting an acceptable threshold for NZ society in high stake applications, such as healthcare, is a good example of the need for input from policymakers and a feedback loop among all experts.

# 3. Case Study

## 3.1. Data

For this simplified case study, we use a small subset of the New Zealand General Practice (GP) Electronic Health Record (EHR) data from [19]. The complete data includes three years' medical records for over 9,000 patients from 44 different GP practices across NZ, collected using a stratified random sampling method to minimise data collection bias. The data includes various categories of free-text and tabular data, and were manually processed, annotated and verified by eight GP researchers (For further details of the NZ-GP Harms data, see [19]).

The small section of the above data, referred to as NZ-GP-small data, includes patients from only Urban locations and only the privileged ethical group 'NZ Europeans' and the protected indigenous group Māori. This results in a total of 3,768 patient data. DI for the NZ-GP-small data is 1.23 for Gender (the protected group is females) and 0.99 for Ethnicity.

## 3.2. Problem definition

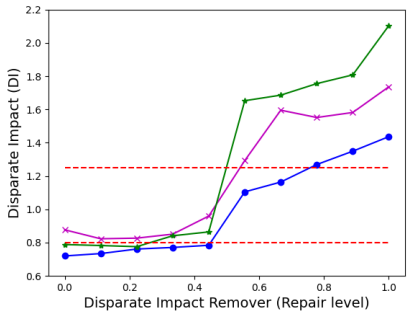
The binary classification task defined for this research is predicting medication-related harm in New Zealand GP from tabular data (see [20] for more details).

## 3.3. Algorithm Selection

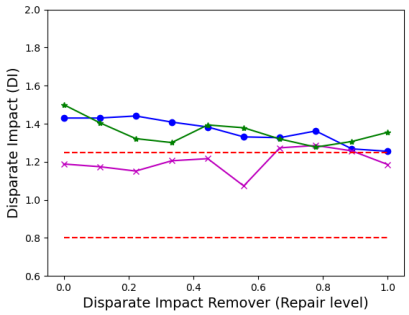
This paper uses logistic regression, XGBoost and random forest as the three machine-learning algorithms for predicting medication-related harm in NZ-GP-small datasets. We use specificity and sensitivity to evaluate the performance of these three algorithms. In addition, we consider the fairness measures using DI, where in accordance with the US regulations, we expect the score to be between 0.8 and 1.25.

Figure 2 provides all three algorithms' DI measure, specificity and sensitivity scores. Firstly, considering DI plots, the fairness is measured based on ethnicity and gender, where protected groups are Māori and female, respectively. In both cases, when the disparate impact remover repair level is at '0', the DI measures of XGBoost are within the red dotted lines, indicating the model fairness is within the acceptable range. However, the logistic regression sensitivity score is the best, and the specificity of all three algorithms is similar.

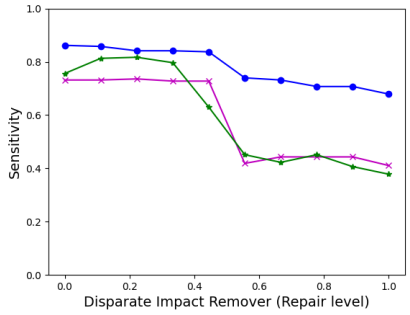
Disparate impact remover [17] is a debias technique used at the preprocessing stage, where it is designed to edit feature values to increase group fairness while preserving rank-ordering within groups. The repair level



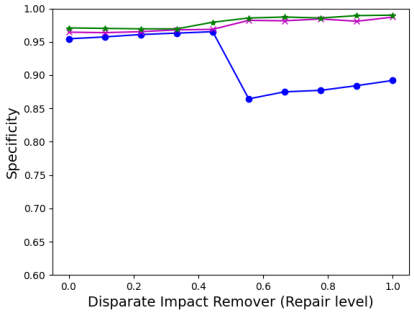

(a) DI measures for ethnicity with protected group Māori




(b) DI measures for gender with protected group Female




(c) Sensitivity scores




(d) Specificity scores

**Figure 2:** Disparate impact (DI) measures for predicting drug harm in NZ-GP-small data for three machine learning models logistic regression, XGBoost and random forest. Disparate impact remover is used as a debiasing technique at the preprocessing stage, increasing repair levels from 0 to 1. The sensitivity and specificity of the models are also presented for direct comparison.

can be set from '0', indicating no change, to 1, indicating maximum repair. In Figure 2a, the increase in repair level varies the DI score, with values 0.4 to 0.6 being the best for all three algorithms. In Figure 2b, the increase in repair level marginally improves the DI values of logistic regression and random forest, with repair values between 0.8 and 0.9 being the best. In contrast, for XGBoost values, 0 to 0.7 is better for the DI scores (see Figure 2b). Unfortunately, the increase in repair levels decreases sensitivity scores for all three classifiers and the specificity score of logistic regression.

### 3.4. Feedback

At each stage, there is a need for a feedback loop, where the decisions impact the final solution. At the data collection and processing stage, clinicians need to discuss the details of the approach with AI experts and ensure that ethical and legal policies are considered. Data bias and mitigation of bias, if any, must also be handled. At the algorithm selection stage, as demonstrated in Section 3.3, the accuracy measures cannot simply determine the choice, but the algorithmic fairness measures must be considered. The feedback loop can vary based on the specification of the problem at hand.

## 4. Discussions and Future Work

This research proposed a development stage AI-based healthcare framework with a feedback loop to ensure a fair real-world practical solution. We argue the need to consider CQI by including feedback at each stage and working together as a team of experts. Recognising the importance of ongoing formal audits and reviews of effect and efficacy in the implementation phase is vital. We use simplified NZ patient data as a case study to demonstrate the framework. We provide early- stage experimental results using machine learning algorithms where it is evident in addition to the accuracy measures, there is also a need to consider fairness measures and mitigation techniques.

As indicated before, the NZ GP dataset is very complex and includes multi-sourced data. Hence, there is a need to extend the process and algorithm selection to enable a fair, practical solution. Furthermore, we only consider ethnicity and gender as binary cases. However, the NZ population and data include other ethnic and minority groups. While we believe the proposed framework allows the flexibility to adopt the simple case presented in this paper to a more complex scenario, it is a much-needed future direction for this research.

## 5. Acknowledgments

VY is supported by the University of Auckland Faculty of Science Research Fellowship program.

## References

[1] C. S. Webster, S. Taylor, C. Thomas, J. M. Weller, Social bias, discrimination and inequity in healthcare: mechanisms, implications and recommendations, BJA Education 22 (2022) 131–137.

[2] V. Yogarajan, Domain-specific language models for multi-label classification of medical text, Ph.D. thesis, The University of Waikato, 2022.

[3] M. A. Ahmad, A. Patel, C. Eckert, V. Kumar, A. Teredesai, Fairness in machine learning for healthcare, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 3529–3530.

[4] B. Giovanola, S. Tiribelli, Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms, AI & Society (2022) 1–15.

[5] S. S. Biswas, Role of Chat GPT in public health, Annals of Biomedical Engineering (2023) 1–2.

[6] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys (CSUR) 54 (2021) 1–35.

[7] V. Yogarajan, G. Dobbie, S. Leitch, T. T. Keegan, J. Bensemann, M. Witbrock, V. Asrani, D. Reith, Data and Model Bias in Artificial Intelligence for Healthcare Applications in New Zealand, Frontiers in Computer Science 4 (2022). URL: https://www.frontiersin.org/articles/10.3389/fcomp.2022.1070493. doi:10.3389/fcomp.2022.1070493.

[8] M. K. Lee, K. Rich, Who is included in human perceptions of AI?: Trust and perceived fairness around healthcare AI and cultural mistrust, in: Proceedings of the 2021 CHI conference on human factors in computing systems, 2021, pp. 1–14.

[9] E. Yoshida, S. Fei, K. Bavuso, C. Lagor, S. Maviglia, The value of monitoring clinical decision support interventions, Applied clinical informatics 9 (2018) 163–173.

[10] J. Feng, R. V. Phillips, I. Malenica, A. Bishara, A. E. Hubbard, L. A. Celi, R. Pirracchio, Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare, NJP Digital Medicine 5 (2022) 66.

[11] C. P. McLaughlin, A. D. Kaluzny, Continuous quality improvement in health care: theory, implementation, and applications, Jones & Bartlett Learning, 2004.

[12] M. Maurer, R. Mangrum, T. Hilliard-Boone, A. Amolegbe, K. L. Carman, L. Forsythe, R. Mosbacher, J. K. Lesch, K. Woodward, Understanding the influence and impact of stakeholder engagement in patient-centered outcomes research: a qualitative study, Journal of General Internal Medicine 37 (2022) 6–13.

[13] M. L. Hudson, K. Russell, The Treaty of Waitangi and research ethics in Aotearoa, Journal of Bioethical Inquiry 6 (2009) 61–68.

[14] L. Esmail, E. Moore, A. Rein, Evaluating patient and stakeholder engagement in research: moving from theory to practice, Journal of Comparative Effectiveness Research 4 (2015) 133–145.

[15] S. Kalkman, J. van Delden, A. Banerjee, B. Tyl, M. Mostert, G. van Thiel, Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence, Journal of Medical Ethics 48 (2022) 3–13.

[16] D. Pessach, E. Shmueli, A review on fairness in machine learning, ACM Computing Surveys (CSUR) 55 (2022) 1–44.

[17] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, in: proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data mining, 2015, pp. 259–268.

[18] P. Besse, E. del Barrio, P. Gordaliza, J.-M. Loubes, L. Risser, A survey of bias in machine learning through the prism of statistical parity, The American Statistician 76 (2022) 188–198.

[19] S. Leitch, S. Dovey, W. Cunningham, K. Wallis, K. Eggleton, S. Lillis, A. McMenamin, M. Williamson, D. Reith, A. Samaranayaka, et al., Epidemiology of healthcare harm in New Zealand general practice: a retrospective records review study, BMJ open 11 (2021) e048316.

[20] S. Leitch, S. M. Dovey, W. K. Cunningham, A. J. Smith, J. Zeng, D. M. Reith, K. A. Wallis, K. S. Eggleton, A. W. McMenamin, M. I. Williamson, et al., Medication-related harm in New Zealand general practice: a retrospective records review, British Journal of General Practice 71 (2021) e626–e633.