12-1-2025

Designing with, Not For: Addressing AI Bias through Community-Led Co-Design in Heart Failure Care

Claris Chung
University of Canterbury, claris.chung@canterbury.ac.nz

Sandra Hanchard
University of Auckland, sandra.hanchard@auckland.ac.nz

Yuming Li
Auckland University of Technology, yuming.li@aut.ac.nz

Yvonne Hong
Victoria University of Wellington, yvonne.hong@vuw.ac.nz

Designing with, Not For: Addressing AI Bias through Community-Led Co-Design in Heart Failure Care

Full research paper

Claris Chung
Accounting and Information Systems
University of Canterbury
Christchurch, New Zealand
Email: claris.chung@canterbury.ac.nz

Sandra Hanchard
Department of General Practice and Primary Health Care
University of Auckland
Auckland, New Zealand
Email: sandra.hanchard@auckland.ac.nz

Yuming Li
Computer and Mathematical Sciences
Auckland University of Technology
Auckland, New Zealand
Email: yuming.li@aut.ac.nz

Yvonne Hong
School of Information Management
Victoria University of Wellington
Wellington, New Zealand
Email: yvonne.hong@vuw.ac.nz

Abstract

Artificial Intelligence (AI) for healthcare holds immense promise but carries a profound risk of amplifying existing health inequities, particularly for underserved groups like Pacific peoples in New Zealand. Standard AI models can perpetuate and scale social and clinical biases that lead to poorer health outcomes. This paper argues that to build equitable AI, we must move beyond purely technical fixes and adopt a new methodology grounded in community partnership. We propose an Equity-Centred Co-Design Framework that directly targets the sources of bias. Using the development of an AI-powered management system for Pacific heart failure patients as a case study, we demonstrate how this framework is applied. By integrating Pacific worldviews, our approach ensures that the community's lived experience shapes the AI's development from its foundation. This paper offers a practical roadmap for researchers and developers to create AI systems that are trustworthy, culturally responsive, and grounded in social justice principles.

Keywords
Artificial Intelligence, Co-design, Health Equity, Algorithmic Bias, Pacific Health, Heart Failure

1 Introduction

The development of AI in cardiovascular medicine demonstrated the potential of a paradigm shift in population health management, clinical decisions, and patient outcomes (Meder et al. 2025). AI-driven systems can analyse routine electrocardiogram (ECG) images to identify individuals at high risk of developing heart failure (HF) before the onset of symptoms (Eisen 2025). In New Zealand, AI is being used to create virtual 3D heart models to improve the success rates of ablation procedures, a direct application aimed at preventing HF progression (National Heart Foundation of New Zealand 2021). For clinicians, this blend of predictive insight and procedural precision empowers them to shift from simply treating disease to actively preventing it, allowing for earlier and more effective interventions. Therefore, this could mean fewer hospital stays and more lives saved from this condition for patients and their families.

However, this technological promise can be jeopardised by a risk of perpetuating and amplifying the profound health inequities already faced by underserved communities, such as Pacific and Māori populations (Yogarajan et al. 2022). In Aotearoa New Zealand, Pacific and Māori populations experience a disproportionate burden of cardiovascular disease, including higher incidence rates, earlier onset, and poorer health outcomes compared to other ethnic groups (Chan et al. 2024). Despite the enormous academic and industry investment in healthcare AI, systemic bias can be formed if the AI models are trained with unrepresentative data (Aquino 2023). Therefore, the AI systems can end up as tools that actively harm underserved communities by encoding, scaling, and legitimising existing structural biases under a veneer of objective, data-driven science (Meder et al. 2025).

This can become deeply problematic for Pacific HF patients, as their symptoms are often misattributed or unrecognised due to their unrepresentative data, with factors including systemic racism, leading to avoidable delays in care. For example, in a qualitative study of Pacific HF patients, participants reported presenting to the hospital feeling distressed with symptoms such as shortness of breath, yet were sent home without screening for a heart condition (Hanchard et al. 2024). Another significant barrier to timely care for Pacific patients stems from physiological differences that affect standard diagnostic tests. The biomarker NT-proBNP, which rises during a heart failure (HF) event, is naturally lower in Pacific peoples than in other ethnic groups. Consequently, a Pacific patient experiencing heart failure might not show NT-proBNP levels high enough to meet the typical diagnostic criteria, causing their condition to be overlooked or diagnosed late (Pearson et al. 2025). Without a radical and fundamental shift in the AI development methodology, models are developed to be blind to their unique physiological, cultural, and social realities.

This process risks embedding inequity more deeply into the healthcare infrastructure, making it harder to identify and challenge because it is masked by the perceived objectivity of the algorithm (McCradden et al. 2020). The promise of AI can only be realised if the perils of algorithmic bias are confronted not as a technical glitch to be patched, but as a core socio-technical challenge requiring a new, equity-focused paradigm for development. Therefore, we believe the AI systems should be designed with communities whose voices are often unheard of and who carry disproportionate health and social burdens compared to other groups in our society.

In this context, co-design should not be regarded as a peripheral addition to AI development practices; rather, it constitutes the central mechanism for advancing equity. As a participatory methodology, co-design reorients the development process from designing *for* communities to designing *with* them, thereby ensuring that underserved populations' perspectives and distinct needs are embedded within the very foundations of AI systems. This paper, therefore, introduces an Equity-Centred Co-Design Framework that directly engages with and addresses the structural sources of bias. Drawing on the development of the Pacific-centred Heart Failure Management System (PACE-HF) as a case study, we present a practical roadmap for researchers, illustrating how this framework can be operationalised to create AI systems that are not only technologically robust but also trustworthy, culturally responsive, and fundamentally just.

2 Background and Related Work

2.1 A Crisis in Heart Failure Equity

The burden of heart disease in New Zealand is not shared equally. For Māori and Pacific peoples, heart failure is not merely a chronic condition but a profound crisis defined by staggering disparities in prevalence, premature onset, and mortality (Chan et al. 2024). In 2020, the mortality rate from heart disease for Māori was 99.6 deaths per 100,000, more than double the rate for European/Other

populations. For Pacific peoples, cardiovascular disease (CVD) is the leading cause of death, responsible for one in every three Pacific deaths (National Heart Foundation of New Zealand n.d.). Between 2006 and 2018, while heart failure incidence rates for individuals aged 70 and over declined for Europeans, it remained static for Māori and Pacific peoples (Chan et al. 2024). This indicates that the current healthcare system, even as it improves on average, is structured in a way that disproportionately benefits the European population, thereby actively widening the relative gap.

2.2 The Double-Edged Sword of AI in Cardiology

These persistent inequities are not accidental but are symptoms of deeply embedded biases operating at multiple levels of the healthcare journey. The appeal of AI in cardiology is undeniable and well-founded, and it can even be an ultimate solution for the disease and inequity in healthcare. AI-powered algorithms have demonstrated their capacity to transform cardiovascular practice across multiple domains. For example, AI has proven superior to human sonographers in assessing cardiac function from echocardiogram images, leading to more accurate and efficient interpretations (He et al. 2023). Beyond diagnostics, AI can identify novel biomarkers and complex patterns invisible to human clinicians, moving beyond traditional risk scores to offer more precise, individualised prognostication by processing vast, multimodal datasets (Biondi-Zoccai et al. 2025).

Despite this promise, AI can carry a profound risk rooted in a simple principle: "garbage in, garbage out." In the context of healthcare, this translates to "bias in, bias out" (Yale School of Medicine 2024). If the training data is not a fair and accurate representation of the full diversity of the human population on which the final tool will be deployed, the resulting model will inevitably exhibit performance deficits for the underrepresented groups (Norori et al. 2021). In healthcare, this problem is pervasive. The vast datasets required to train complex AI models are typically sourced from a small number of large, academic medical centres located in affluent, urban areas within a few high-income countries (Stetler 2024). Consequently, these datasets are heavily skewed towards patients of European ancestry and often lack sufficient representation of Indigenous peoples, ethnic minorities, and populations from different geographic or socioeconomic backgrounds (Chinta 2024). A systematic review of AI-based risk prediction models for cardiovascular disease starkly illustrates this point. The analysis found that a majority (60%) of the models were developed using data from the United States, and, critically, only 32% of the studies even reported the race and ethnicity of the participants in their datasets (Provost et al. 2025).

2.3 The Triad of Bias: A Pernicious Cycle

As such, if that data reflects the historical biases and inequalities of the society and systems that produced it, the AI will learn, replicate, and often amplify those same biases (Norori et al. 2021). Aquino (2023) framed this problem as a "pernicious cycle" involving three interconnected types of bias: social, clinician, and technological. Social bias encompasses the systemic barriers and societal prejudices that shape a patient's health journey long before they enter a clinic. Clinician bias involves the cognitive and environmental factors that influence a provider's decision-making, from diagnostic uncertainty to implicit assumptions. Finally, technological bias emerges when digital tools and AI models, trained on data reflecting these existing inequities, encode and amplify them under a layer of objectivity. These three are not isolated; they form a feedback loop where societal inequities shape clinical interactions, which in turn generate the biased data used to build flawed technologies.

2.4 Co-Design as a Methodological Imperative

Therefore, the naïve approach of traditional AI development, which often excludes user voices, is insufficient and poses an ethical risk in a high-stakes field like healthcare. To address this, a methodological shift is required. In the health system of Aotearoa New Zealand, co-design based on community values is well-recognised as a critical methodology for achieving health equity in service design (Te Tāhū Hauora Health Quality & Safety Commission 2024). Co-design, a participatory approach that embeds user voices into the development process, offers a powerful way to confront these biases at their source. An equitable future for medical AI requires a new paradigm grounded in co-design principles, authentic community partnership, and the recognition of Indigenous data sovereignty. Co-design intentionally brings together all relevant stakeholders from the very beginning of a project, and the goal is to collaboratively understand experiences, define problems, and generate solutions together. This approach is not simply a matter of good practice; for Māori, lifting whānau voices in health research is an obligation for honouring the partnership principles of Te Tiriti o Waitangi (the Treaty of Waitangi), the founding document of Aotearoa New Zealand (Goodwin 2024). For both Māori and Pacific peoples, co-design is a vital mechanism for redressing long-standing health inequities. OL@-OR@ project involved the co-design of a mobile health (mHealth) program aimed at supporting healthy lifestyles for

Māori and Pasifika communities. The process involved extensive community engagement to ensure the app and website were culturally relevant and met the specific needs of the users (Chinta 2024).

2.5 Research Question and Objectives

This review establishes that while AI holds promise, its uncritical application risks harming the very communities that stand to benefit most. It also posits that co-design is the necessary methodological approach to mitigate this risk. However, a gap remains in understanding how to practically integrate co-design into the technical lifecycle of AI development to systematically address bias. This leads to our central research question: How can a co-design methodology be structured within the AI development process to directly confront and mitigate the triad of social, clinician, and technological bias?

To answer this, our objective is to propose and detail a practical framework that maps specific co-design activities to each phase of the Design Science Research (DSR) process, using the development of an AI-powered heart failure management system for Pacific peoples as an illustrative case study.

3 Research Design and Methodology: An Equity-Centred Co-Design Framework

The development of the PACE-HF system is guided by the Equity-Centred Co-Design Framework that intentionally weaves together Pacific worldviews with rigorous information systems development theory. Our methodology uses Nunamaker's multimethod framework for Design Science Research (DSR) (Nunamaker et al. 1990) to provide the structured, iterative steps for system design, development, and evaluation. The Tongan Kakala (garland) research framework (Fua 2014; Helu-Thaman 2007) provides the guiding cultural principles for each phase, ensuring the process is relational and respectful. Importantly, co-design is not just one of these steps; it is the specific, participatory methodology we employ to execute each phase of the DSR cycle. Through this integrated approach, our 5-phase project collects and integrates small and big data to develop an AI model that delivers personalised, culturally grounded, post-discharge support for Pacific patients with a heart failure condition.

Table 1 summarises our Equity-Centred Co-Design Framework in action. It details how each of the five project phases is guided by a specific Pacific principle, executed through a corresponding DSR step and co-design methodology, and defined by its required inputs and expected outcomes.

3.1 Phase 1: Teu – Grounding the Design in Community

The Kakala framework begins with Teu, the careful preparation, conceptualising, and planning required to create a beautiful garland (Fua 2014). This principle mandates that we begin by establishing respectful relationships and a shared understanding, which aligns with the DSR phase of Observation & Theory Building. We execute this phase through an Experience-Based Co-Design (EBCD) approach (Van Citters 2017). Community workshops, respectful dialogue, and collaborative sense-making generate conceptual foundations by defining the problem, requirements, and culturally grounded definitions of success. This builds the "theory" base from lived experiences and cultural knowledge. Also, our study is co-led by domain experts and researchers who can serve as trusted voices of the community. For example, the leadership team includes a cardiovascular equity researcher whose work has been translated into practice through collaborations with Pacific health leaders, alongside a Tongan researcher with extensive experience leading qualitative research with diverse Pacific communities in a heart health equity programme.

In these workshops, we collaboratively create journey maps and gather patient narratives to identify the emotional and practical "touchpoints" of the heart failure journey. This process is fueled by the lived experiences of Pacific peoples, the clinical expertise of providers, and the cultural knowledge of community advisors. The expected cultural outcome is a trusting and safe research environment with a co-created definition of success, while the technological outcome is a rich qualitative dataset of culturally resonant HF symptom expressions and a community-validated design brief.

3.2 Phase 2: Toli (Part 1) – Co-Governing Population-Level Knowledge

The next step is Toli, the act of picking the flowers for the garland. The metaphor is important as this principle guides our data collection, ensuring it is done with care and respect. The first part of Toli involves gathering the broad, existing knowledge needed for the garland (Helu-Thaman 2007). This corresponds to the DSR phase of System Development & Observation. The primary co-design activity here is the oversight provided by the Pacific-led ethics and equity advisory board on the collection and governance of population-level health "big data" that will be used to develop the technical infrastructure

for AI models. The Pacific-led advisory board co-governs the process of accessing and using de-identified "big data" (e.g., hospitalisation records), ensuring its use aligns with the community's values and the principles of Pacific Data Sovereignty (Ministry for Pacific Peoples 2023).

Kakala Framework	DSR & Co-Design	Required Inputs	Expected Outcomes
Phase 1: Teu – Grounding the Design in Community	Observation & Theory Building Experience-Based Co-Design	Lived experiences of Pacific peoples; clinical expertise of providers; cultural knowledge of community advisors	**Cultural**: Trusting, safe research environment with a co-created definition of success. **Technological**: Rich qualitative dataset of culturally resonant HF symptom expressions; community-validated design brief.
Phase 2: Toli (Part 1) – Co-Governing Population-Level Knowledge	System Development & Observation Co-governance of big data use in alignment with Pacific values	Ethics-approved access to hospitalisation records, CVD risk factor database; oversight from Pacific-led ethics & equity advisory board	**Cultural**: Transparent, accountable process for population data use; community trust upheld. **Technological**: Foundational AI models for baseline risk stratification and deterioration prediction.
Phase 3: Toli (Part 2) – Co-Creating Personal Stories	System Development Small Data Acquisition via Participatory Data Collection	Functioning prototype system; active, consented participation of Pacific peoples	**Cultural**: Empowering data collection experience where participants feel agency over personal information. **Technological**: Creation of a unique, culturally-specific “small dataset” for AI fine-tuning.
Phase 4: Tui – Weaving Knowledge through Iterative Co-Evaluation	Experimentation & Observation & System Development Model Integration & System Refinement, via Iterative Co-Evaluation	Integrated AI models; evolving prototypes; continuous feedback from community & clinical partners	**Cultural**: System reflects feedback, strengthening trust and ownership. **Technological**: Final integrated hybrid AI model with refined UI and Explainable AI (XAI) functionality.
Phase 5: Luva & Malie/Mafana – Evaluation & Gifting Back	Observation & Theory Building Summative Co-Evaluation	Integrated PACE-HF system; active participation of pilot cohort	**Cultural**: Evidence of system trust and acceptability; honoured in a celebratory gifting-back event. **Technological**: Quantitative usability data; robust evaluation of AI accuracy, fairness, and trustworthiness.

Table 1. Equity-Centred Co-Design Framework

This process is executed by accessing a range of health big data, including hospitalisation records and the CVD risk factor database (e.g., PREDICT in New Zealand), all guided by the active governance of the advisory board. Then, culturally, it ensures a transparent and accountable process for using population data that upholds community trust. Technologically, it will deliver the foundational, population-level AI models for baseline risk stratification and deterioration prediction.

3.3 Phase 3: Toli (Part 2) – Co-Creating Personal Stories

The second part of Toli involves carefully selecting the specific, beautiful flowers that will make the garland personal and unique (Helu-Thaman 2007). This aligns with the DSR phase of System Development (Small Data Acquisition), executed through Participatory Data Collection. Participants are active partners in co-creating the "small data" (symptom inputs, wearable data, voice notes) that will be used to personalise the AI. This process honours their lived experience as a precious "flower" that is essential for the garland's final form.

To facilitate this participatory process, the required inputs are a functioning prototype system and the active, consented participation of Pacific peoples. The expected cultural outcome is an empowering data collection experience where participants feel agency over their personal information, while the technological outcome is the creation of a unique, culturally-specific "small dataset" essential for fine-tuning the AI.

3.4 Phase 4: Tui – Weaving Knowledge through Iterative Co-Evaluation

Tui is the act of stringing the flowers and weaving the garland, representing analysis and integration (Helu-Thaman 2007). This principle guides us to respectfully weave together the different forms of knowledge – community, clinical, and data-driven – into a coherent whole. This aligns with the DSR phase of Experimentation & Observation & System Development(Model Integration and System Refinement), which is driven by Iterative Co-Evaluation. We integrate the "big data" and "small data" models and then engage community and clinical partners in rapid, participatory feedback cycles. Using techniques like "think-aloud" usability testing with evolving prototypes (Jacob et al. 2025), their feedback directly and immediately shapes the next iteration of the system. This act of "weaving" ensures the final system is a synthesis of all the knowledge gathered.

This weaving process is fuelled by three key inputs: the integrated AI models, evolving prototypes, and the continuous feedback from our community and clinical partners. The expected cultural outcome is a system that demonstrably reflects this feedback, strengthening trust and ownership. In contrast, the technological outcome is the final, integrated hybrid AI model, featuring a refined user interface and transparent Explainable AI (XAI) functionality.

3.5 Phase 5: Luva & Malie/Mafana – Evaluation & Gifting Back

The framework culminates in Luva (evaluation) and Malie/Mafana (gifting back with warmth). These principles ensure our work is not only validated by the community but also results in a tangible benefit that is shared and celebrated. This corresponds to the DSR Demonstration and Evaluation phase (via Observation & Theory Building), implemented as a Summative Co-Evaluation. The pilot study serves as the formal Luva. The process concludes with the Malie/Mafana principle, where the final system (the "garland") is gifted back to the community in a celebratory event, closing the loop of reciprocity and shared success.

To bring this final phase to fruition, the required inputs are the integrated PACE-HF system and the active participation of the pilot cohort. The process is designed to yield significant cultural outcomes, including evidence of system trust and acceptability, which will be honoured at a celebratory event. Technologically, this phase will deliver quantitative data on usability and a robust evaluation of the AI's accuracy, fairness, and trustworthiness.

4 Enabling Technologies for Equitable AI

Our co-design methodology, which centres the lived experiences of Pacific peoples, revealed two core requirements for an equitable AI system. First, it must integrate evidence-based clinical knowledge with a deep cultural and contextual understanding, including the social determinants of health (SDOH) that are particularly relevant for communities carrying high social burdens. Second, the system must be transparent to build the trust necessary for adoption (Consoli et al. 2025; Tayal 2025). To meet these community-defined needs, our technical framework is built on two pillars: a Dual Large Language Model (Dual-LLM) Architecture and a deep commitment to Explainable AI (XAI).

The Dual-LLM architecture creates a system that embodies a form of bicultural competence by weaving together quantitative clinical evidence with qualitative cultural understanding. The integration of XAI tools renders the AI's decision-making processes transparent, empowering both clinicians, who can validate its reasoning, and patients, who can understand how their unique context has been considered.

4.1 A Dual-LLM Hybrid Architecture

To address the insights from our co-design process and dismantle the feedback loop of algorithmic bias, we propose a Dual-LLM architecture designed to integrate two distinct but complementary forms of knowledge: the vast, quantitative world of biomedical evidence and the deep, qualitative world of lived experience and cultural context, and the social determinants of health that shape patient outcomes (Consoli et al. 2025). This hybrid approach represents a deliberate shift away from a single, decontextualised model toward a synergistic system that enhances culturally-responsive decision-making by clinicians.

To ensure seamless integration, we employ a chained architecture where outputs from the Generalist LLM (LLM-G) are fed as prompts into the Community LLM (LLM-C), allowing for iterative refinement of predictions with cultural nuance. This reduces inconsistencies and enhances computational efficiency, drawing on recent advancements in modular LLM designs for healthcare applications (Maity and Saikia 2025) .

Furthermore, inspired by Mixture-of-Experts (MoE) extensions in dual setups, we incorporate dynamic routing between LLMs based on input type, optimising for equity in underrepresented data scenarios (Raschka 2025)

4.1.1 The Generalist Engine (LLM-G): Foundations in "Big Clinical Data"

The first component of the architecture is a large, "Generalist" Large Language Model (LLM-G). This model serves as the system's foundation, providing a robust and comprehensive understanding of general medical knowledge. The development of LLM-G involves pre-training on massive, multimodal clinical datasets (Choi et al. 2017). These datasets encompass the full breadth of biomedical information, including:

- Medical Literature: Vast amounts of text from sources like PubMed abstracts and full-text research articles, enabling the model to learn from the global body of scientific evidence.
- Electronic Health Records (EHRs): Large-scale, de-identified EHRs provide real-world data on disease presentation, clinical workflows, and treatment patterns. In New Zealand, this "big data" is available through the Health Data Platform, with which our team has a close working relationship (University of Auckland n.d.).
- Clinical Guidelines and Textbooks: These sources provide the model with structured, evidence-based knowledge on diagnostic criteria and standard-of-care treatment protocols.
- Multimodal Data: Integration of data beyond text, such as imaging reports, laboratory results, and electrocardiogram (ECG) data, allows the model to synthesise information from disparate sources to form a holistic clinical picture (University of Auckland n.d.).

Through this extensive training, LLM-G develops the capacity to perform a wide range of complex clinical and administrative tasks. It can accurately summarise lengthy patient records, extract structured information from unstructured clinical notes, answer complex medical questions, and assist in clinical decision support by retrieving relevant evidence from the literature. To mitigate inherent biases in big data, we incorporate debiasing techniques during pre-training, such as counterfactual data augmentation and empathy-based reinforcement learning, ensuring better representation of underrepresented groups like Pacific populations (Templin et al. 2025).

4.1.2 The Community Voice (LLM-C): Fine-Tuning on "Small, Rich Data"

The second component, the "Community" LLM (LLM-C), is designed to provide the cultural and contextual nuance that the LLM-G lacks. This is achieved through a process called fine-tuning, where a pre-trained foundation model undergoes additional, specialised training on a smaller, domain-specific dataset. This technique allows the model to adapt its general linguistic capabilities to a specific task, dialect, or knowledge domain without the prohibitive cost of training a new model from scratch (Bootcamp AI 2025) .

For the LLM-C, the fine-tuning dataset is not "big data" but "small, rich data", a multimodal collection of qualitative, narrative, and real-time information gathered directly from and in partnership with the community. This data is curated to capture the lived experiences that are invisible in standard clinical datasets. Key sources for this data include:

- Community Narratives and Knowledge: Transcripts from interviews with patients and families about their health journeys, integrated with principles of mātauranga Māori and Pacific wisdom to ensure recommendations align with holistic models of health (Saunders et al. 2024).
- Patient-Generated Health Data: The framework also incorporates real-time, multimodal data generated directly by patients using the PACE-HF prototype, including:
    - Unstructured Text: Patient-entered symptom descriptions, providing direct, in-the-moment accounts of their health.
    - Audio/Visual Data: Short voice or video symptom diaries, capturing nuances of tone and expression that text cannot convey.
    - Time-Series Data: Objective physiological signals from wearables (e.g., heart rate, weight, activity), providing quantitative context to the patient's subjective narrative.

We create a holistic, longitudinal profile of each patient's health journey by weaving together these different data streams, from community narratives to daily physiological signals. Given the often-limited size of such specialised datasets, specific technical strategies are required. Parameter-Efficient Fine-Tuning (PEFT) methods, such as LoRA, are ideal as they dramatically reduce computational cost by only training a small number of new parameters (Dialzara Team 2025). This makes it feasible to train highly specialised models on small, high-quality datasets. An additional strategy involves using a larger model to generate high-quality synthetic data that reflects the nuances of the small, real-world dataset, which can then be used to fine-tune the LLM-C more robustly.

4.2 Explainable AI (XAI) for Trust and Transparency

For any clinical AI system to be successfully adopted, particularly one designed to serve communities with a history of mistrust in the healthcare system, it must be demonstrably trustworthy. Technical accuracy alone is insufficient; building trust demands transparency to mitigate risks of perpetuating inequities (Sagona et al. 2025). The "black box" nature of many advanced AI models, where the reasoning behind a prediction is opaque even to its developers, represents a significant barrier to clinical integration, exacerbating issues like hallucinations and unreliable outputs in generative healthcare applications (Maity and Saikia 2025). Explainable AI (XAI) addresses this by providing tools and methods to make AI decision-making processes transparent and interpretable, distinguishing between local (instance-level) and global (model-wide) explanations to suit diverse stakeholders (Sadeghi et al. 2024). Within the proposed Dual-LLM framework, XAI is not an optional add-on but an essential component for validating the model's reasoning, identifying hidden biases, and empowering both clinicians and patients in a shared decision-making process.

Among the growing suite of XAI techniques, two model-agnostic methods have become particularly prominent for interpreting complex models: LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) (Vimbi et al. 2024). Given the hybrid nature of the proposed Dual-LLM architecture, an effective XAI strategy must be tailored to each model component's distinct functions and the different end-users' needs. This involves a strategic application of both SHAP and LIME, transforming the system's output from a simple prediction into a transparent, interactive dialogue.

For the Generalist Engine (LLM-G), which is tasked with making quantitative clinical predictions based on structured and unstructured electronic health records (EHR) data, SHAP is the ideal tool. Its mathematical rigour and ability to provide both global and local feature attributions are perfectly suited to the needs of clinicians and researchers. For example, when the LLM-G predicts a high 10-year risk of a cardiovascular event for a patient, a SHAP summary plot can provide a global explanation, showing that across the entire patient population, factors like NT-proBNP levels, systolic blood pressure, and smoking history are the most significant predictors (Lu et al. 2021). More importantly, a local SHAP force plot for that specific patient can show the clinician precisely how their individual values for these features contributed to their high-risk score (González 2022). This allows the clinician to validate the model's reasoning against their own knowledge and to have a more informed conversation with the patient about which specific risk factors need to be addressed. The extensive use of SHAP in heart failure prediction research confirms its utility and robustness in this clinical domain (Lu et al. 2021).

To demonstrate this approach, we first generated a synthetic dataset with clinically plausible distributions of routinely collected features (e.g., NT-proBNP, systolic blood pressure, smoking history), allowing us to illustrate the methodology without relying on patient-identifiable records. We then implemented a machine learning pipeline and trained a tree-based ensemble model (XGBoost, just for

example) to predict the 10-year risk of cardiovascular events, ensuring robust internal validation through stratified train–test splitting. SHAP was subsequently employed to provide both global and local explanations of the model's outputs. At the global level, SHAP summary and bar plots (Figures 1a and 1b) highlighted the relative contribution of key predictors across the population, confirming that NT-proBNP levels, systolic blood pressure, and smoking history were consistently among the most influential risk factors.

Figure 1. SHAP beeswarm summary plot showing the distribution of feature contributions to the 10-year cardiovascular risk prediction across the test population. Each dot represents a patient; colour encodes the feature value (red = high, blue = low). Figure 2. SHAP bar summary plot ranking the mean absolute feature contributions. NT-proBNP, systolic blood pressure, and smoking history emerge as the dominant predictors.

At the local level, we generated SHAP force and waterfall plots for individual patients (Figures 2a and 2b), enabling clinicians to see how patient-specific values of NT-proBNP, blood pressure, and smoking status shifted the predicted risk relative to the model's baseline expectation. These patient-level explanations allowed us to simulate clinical use cases in which a cardiologist could compare the model's reasoning against their own judgment, identify concordance or discrepancies, and communicate more transparently with patients about which modifiable risk factors warrant targeted intervention.

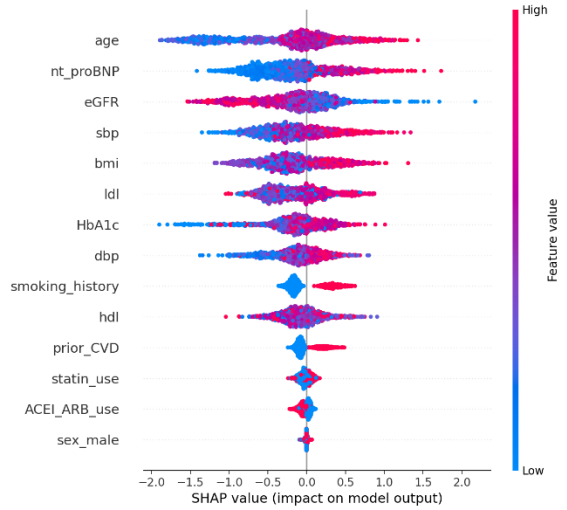

Figure 1. SHAP beeswarm summary plot

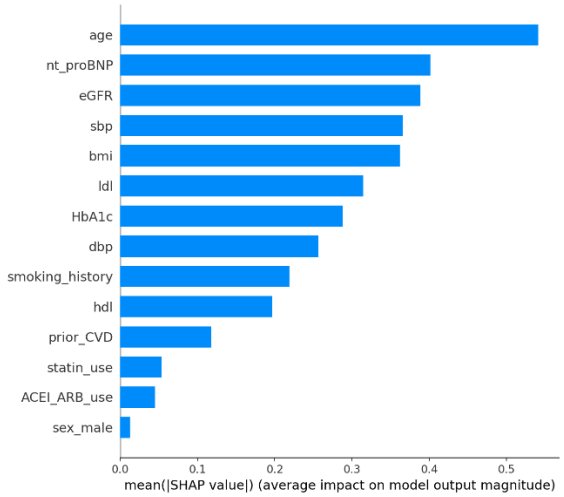

Figure 2. SHAP bar summary plot

Figure 3. SHAP force plot (static rendering) for a representative high-risk patient, indicating how elevated NT-proBNP, increased systolic blood pressure, and smoking history pushed the risk upward (red), while protective factors such as statin use reduced the risk (blue).

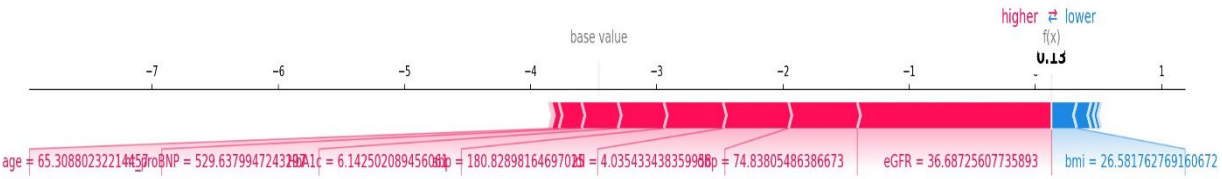

Figure 3. SHAP force plot (static rendering)

Figure 4. SHAP waterfall plot for the same patient, detailing the additive contributions of individual features from the expected baseline risk to the final predicted risk score.

For the Community Voice (LLM-C), which generates culturally nuanced recommendations based on qualitative data, LIME is the more appropriate tool. The goal here is not to quantify the contribution of numerical features, but to understand how specific parts of a narrative or cultural context influenced the model's output. LIME excels at this by highlighting the specific words or phrases in the input text that were most influential in generating a particular recommendation (Sathyan et al. 2022). For instance, if

the system recommends involving whānau in medication management discussions, a LIME explanation could highlight the patient's statement "my children help me remember things" as the key driver for that suggestion. This provides an intuitive, easy-to-understand explanation that builds trust with the patient by showing that their personal story was "heard" and acted upon by the system.

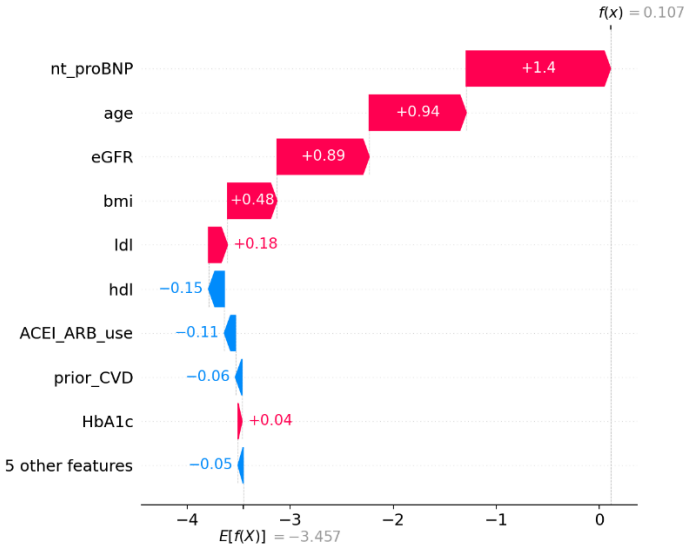

Figure 4. SHAP waterfall plot

To demonstrate this capability, we trained a lightweight text classification model using synthetic but clinically plausible narratives that capture common cultural and contextual factors influencing care. We then applied LIME to a representative case in which the narrative included the statement “my children help me remember things”. The model generated a recommendation to involve whānau in medication management, and the LIME explanation highlighted the reference to children and whānau as the most influential tokens driving this decision (Figure 5).

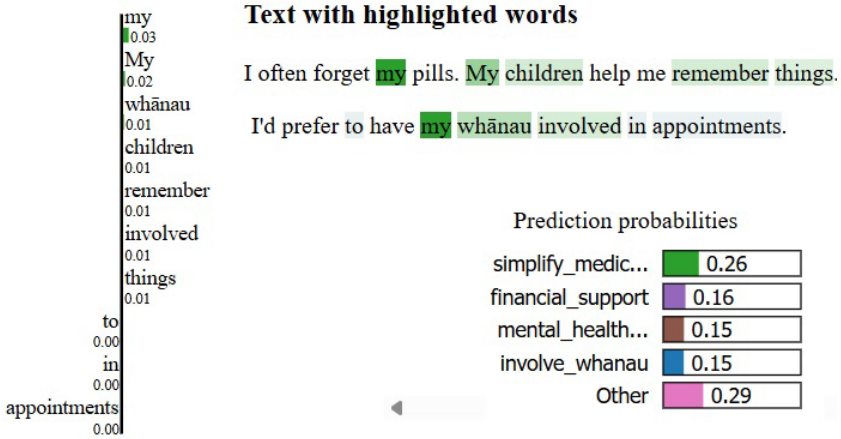

Figure 5. LIME explanation for a narrative-driven recommendation

This example illustrates how narrative-driven recommendations can be made transparent by directly linking model outputs to patient-expressed experiences. Such explanations provide clinicians with a clear rationale for culturally sensitive recommendations and reassure patients that their voices are meaningfully integrated into the decision-making process.

5 Implications, Contributions, and Limitations

The methodology and framework proposed in this paper offer significant contributions to health equity, information systems research, and the ethical development of AI. By treating algorithmic bias as a socio-

technical challenge rather than a purely technical one, our approach provides a new paradigm for creating equitable health technologies.

5.1 Contribution to Health Equity

This project's primary contribution is a direct intervention to reduce the profound health disparities in heart failure faced by Pacific peoples. The co-design process, grounded in Pacific worldviews, ensures that the PACE-HF system is not just another piece of technology, but a culturally-grounded tool designed to address specific, community-identified needs. By creating a system that understands culturally resonant expressions of symptoms and is built on a foundation of trust, we aim to reduce delays in care, improve post-discharge management, and empower patients and their families to become more confident partners in their own healthcare. This work provides a tangible pathway to move beyond simply describing health inequities to actively dismantling them through community-led technological innovation.

5.2 Contribution to Information Systems/AI Research

Methodologically, this paper contributes a novel framework that integrates Indigenous research principles (the Kakala framework) with the established structure of Design Science Research (DSR). We demonstrate how co-design can be operationalised as the core methodology executing each phase of the DSR cycle, offering a practical roadmap for other researchers.

Technologically, our proposed dual-data, dual-LLM architecture presents an innovative solution to the pervasive problem of unrepresentative data. By synergising a "Generalist" model trained on big clinical data with a "Community" model fine-tuned on small, rich, culturally-specific data, we offer a new approach to building AI that is both clinically robust and contextually aware. Furthermore, our strategic use of Explainable AI (XAI) contributes to the growing body of research on how to build transparent and trustworthy systems, which is a critical prerequisite for clinical adoption.

5.3 Ethical Contributions

This framework is fundamentally grounded in an ethics-first approach. By embedding the principles of Pacific worldviews into our methodology from the outset, we ensure that the community has ownership and control over their data. Our co-design process, which includes the formation of a Pacific-led advisory board and the co-creation of data governance protocols, provides a practical example of how to move beyond extractive research models towards genuine, power-sharing partnerships. This commitment to a culturally safe and ethically robust process is essential for redressing historical mistrust and ensuring that the development of AI serves, rather than harms, the community.

5.4 Limitation

While this study lays the conceptual and methodological foundation for equitable AI, further work is needed to test its scalability and generalisability beyond Pacific contexts. Deep co-design is resource-intensive, and refining practical pathways for sustainable implementation will be essential. The current dual-LLM model has been tested using synthetic data, and future research will extend validation to real-world clinical environments and longitudinal evaluation. These next steps will strengthen the evidence base and support translation of the framework into broader Indigenous and global health settings.

6 Conclusion

A fundamental shift in methodology is required to realise the promise of AI in healthcare and avoid its perils. This paper has proposed an Equity-Centred Co-Design Framework that places community partnership, cultural values, and shared power at the heart of the AI development process. Through our case study of the PACE-HF system, we have demonstrated how this approach can be used to directly address the technological, social, and clinician biases that perpetuate health inequities.

The innovative combination of deep community co-design, guided by the Tongan Kakala framework, with a sophisticated dual-LLM architecture, offers a new paradigm for development. Co-design is the essential bridge between community, clinic, and code. By following this framework, we can move towards a future where AI becomes a powerful tool for healing, empowerment, and health justice, transforming the heart failure journey for Pacific peoples and offering a blueprint for equitable AI development for all.

7 References

Aquino, Y. S. J. 2023. "Making Decisions: Bias in Artificial Intelligence and Data-Driven Diagnostic Tools," *Australian Journal of General Practice* (52), pp. 439-442.

Biondi-Zoccai, G., D'Ascenzo, F., Giordano, S., Mirzoyev, U., Erol, Ç., Cenciarelli, S., Leone, P., and Versaci, F. 2025. "Artificial Intelligence in Cardiology: General Perspectives and Focus on Interventional Cardiology," *Anatol J Cardiol* (29:4), pp. 152-163.

Bootcamp AI. 2025. "Fine-Tuning Llms: A Guide to Optimal Data Labeling." from https://bootcampai.medium.com/fine-tuning-llms-a-guide-to-optimal-data-labeling-b06c5837e6e7

Chan, D. Z., Grey, C., Doughty, R. N., Lund, M., Lee, M. A. W., Poppe, K., Harwood, M., and Kerr, A. 2024. "Widening Ethnic Inequities in Heart Failure Incidence in New Zealand," *Heart* (110:4), pp. 281-289.

Chinta, S. V. W., Zichong; Palikhe, Avash; Zhang, Xingyu; Kashif, Ayesha; Smith, Monique Antoinette; Liu, Jun; Zhang, Wenbin. 2024. "Ai-Driven Healthcare: A Review on Ensuring Fairness and Mitigating Bias."

Choi, E., Schuetz, A., Stewart, W. F., and Sun, J. 2017. "Using Recurrent Neural Network Models for Early Detection of Heart Failure Onset," *J Am Med Inform Assoc* (24:2), pp. 361-370.

Consoli, B., Wang, H., Wu, X., Wang, S., Zhao, X., Wang, Y., Rousseau, J., Hartvigsen, T., Shen, L., Wu, H., Peng, Y., Long, Q., Chen, T., and Ding, Y. 2025. "Sdoh-Gpt: Using Large Language Models to Extract Social Determinants of Health," *Journal of the American Medical Informatics Association*).

Dialzara Team. 2025. "Fine-Tuning Llms with Small Data: Guide." from https://dialzara.com/blog/fine-tuning-llms-with-small-data-guide

Eisen, J. 2025. "New Ai Tool Identifies Risk of Future Heart Failure." Retrieved 19/August, 2025, from https://medicine.yale.edu/news-article/new-ai-tool-identifies-risk-of-future-heart-failure/

Fua, S. u. J. 2014. "Kakala Research Framework: A Garland in Celebration of a Decade of Rethinking Education." USP Press.

González, S. H., Wan-Ting; Burba, Davide; Chen, Trista Pei-Chun; Wang, Chun-Li; Wu, Victor Chien-Chia; Chang, Shang-Hung. 2022. "Interpretable Estimation of the Risk of Heart Failure Hospitalization from a 30-Second Electrocardiogram."

Goodwin, D. B., Amohia. 2024. "Co-Designing Health Research in Aotearoa New Zealand: Lessons from the Healthier Lives National Science Challenge." Healthier Lives–He Oranga Hauora National Science Challenge.

Hanchard, S., Brewer, K. M., Taueetia-Su'a, T., Vaka, S., Ameratunga, S., Tane, T., Newport, R., Selak, V., Harwood, M., and Grey, C. 2024. "Navigating the Long Journey of Heart Failure-Experiences of Māori and Pacific Peoples," *The New Zealand Medical Journal* (137:1603), pp. 25-32.

He, B., Kwan, A. C., Cho, J. H., Yuan, N., Pollick, C., Shiota, T., Ebinger, J., Bello, N. A., Wei, J., Josan, K., Duffy, G., Jujjavarapu, M., Siegel, R., Cheng, S., Zou, J. Y., and Ouyang, D. 2023. "Blinded, Randomized Trial of Sonographer Versus Ai Cardiac Function Assessment," *Nature* (616:7957), pp. 520-524.

Helu-Thaman, K. 2007. "Kakala: A Pacific Concept of Teaching and Learning [Conference Paper]," *The Australian College of Educators National Conference, Cairns, Queensland*.

Jacob, C., Müller, R., Schüler, S., Rey, A., Rey, G., Armenian, B., Vonlaufen, A., Drepper, M., and Zimmerli, M. 2025. "Think-Aloud Testing of a Companion App for Colonoscopy Examinations: Usability Study," *JMIR Hum Factors* (12), p. e67043.

Lu, S., Chen, R., Wei, W., Belovsky, M., and Lu, X. 2021. "Understanding Heart Failure Patients Ehr Clinical Features Via Shap Interpretation of Tree-Based Machine Learning Model Predictions," *AMIA Annu Symp Proc* (2021), pp. 813-822.

Maity, S., and Saikia, M. J. 2025. "Large Language Models in Healthcare and Medical Applications: A Review," *Bioengineering (Basel)* (12:6).

McCradden, M. D., Joshi, S., Mazwi, M., and Anderson, J. A. 2020. "Ethical Limitations of Algorithmic Fairness Solutions in Health Care Machine Learning," *The Lancet Digital Health* (2:5), pp. e221-e223.

Meder, B., Asselbergs, F. W., and Ashley, E. 2025. "Artificial Intelligence to Improve Cardiovascular Population Health," *Eur Heart J* (46:20), pp. 1907-1916.

Ministry for Pacific Peoples. 2023. "Improving Pacific Data Equity: Opportunities to Enhance the Future of Pacific Wellbeing," M.f.P. Peoples (ed.). Ministry for Pacific Peoples.

National Heart Foundation of New Zealand. 2021. "Artificial Intelligence Could Improve Heart Treatment." from https://www.heartfoundation.org.nz/about-us/news/stories/artificial-intelligence-could-improve-heart-treatment/

National Heart Foundation of New Zealand. n.d. "Pacific Heart Health Statistics." Retrieved 20/Aug, 2025, from https://www.heartfoundation.org.nz/your-heart/pacific-heartbeat/pacific-heart-health-statistics

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., and Tzovara, A. 2021. "Addressing Bias in Big Data and Ai for Health Care: A Call for Open Science," *Patterns (NY)* (2:10), p. 100347.

Nunamaker, J. F., Chen, M., and Purdin, T. D. 1990. "Systems Development in Information Systems Research," *Journal of management information systems* (7:3), pp. 89-106.

Pearson, A. G., Pearson, J. F., Lewis, L. K., Fa'atoese, A., Poppe, K. K., Pemberton, C., Devlin, G., Lund, M., Richards, A. M., and Troughton, R. 2025. "Lower Nt-Probnp Plasma Concentrations in Pacific Peoples with Heart Failure," *ESC Heart Failure*).

Provost, C., Broughan, J., McCombe, G., Kelly, M. O., Ledwidge, M., Cullen, W., and Gallagher, J. 2025. "Artificial Intelligence (Ai) Models for Cardiovascular Disease Risk Prediction in Primary and Ambulatory Care: A Scoping Review," *medRxiv*), p. 2025.2003.2021.25324379.

Raschka, S. 2025. "The Big Llm Architecture Comparison: From Deepseek-V3 to Kimi K2 — a Look at Modern Llm Architecture Design." Retrieved 19/Jul, 2025, from https://magazine.sebastianraschka.com/p/the-big-llm-architecture-comparison

Sadeghi, Z., Alizadehsani, R., Cifci, M. A., Kausar, S., Rehman, R., Mahanta, P., Bora, P. K., Almasri, A., Alkhawaldeh, R. S., Hussain, S., Alatas, B., Shoeibi, A., Moosaei, H., Hladík, M., Nahavandi, S., and Pardalos, P. M. 2024. "A Review of Explainable Artificial Intelligence in Healthcare," *Computers and Electrical Engineering* (118), p. 109370.

Sagona, M., Dai, T., Macis, M., and Darden, M. 2025. "Trust in Ai-Assisted Health Systems and Ai's Trust in Humans," *npj Health Systems* (2:1), p. 10.

Sathyan, A., Weinberg, A. I., and Cohen, K. 2022. "Interpretable Ai for Bio-Medical Applications," *Complex Eng Syst* (2:4).

Saunders, C., Dalziel, P., Reid, J., and McCallum, A. 2024. "Knowledge, Mātauranga and Science: Reflective Learning from the Interface," *Journal of the Royal Society of New Zealand* (54:2), pp. 207-228.

Stetler, C. 2024. "Ai Algorithms Used in Healthcare Can Perpetuate Bias." Retrieved 22/Aug, 2025, from https://www.newark.rutgers.edu/news/ai-algorithms-used-healthcare-can-perpetuate-bias

Tayal, A. S., Devika; Di Eugenio, Barbara; Allen-Meares, Paula G.; Abril, Eulalia P.; Garcia-Bedoya, Olga; Dickens, Carolyn A.; Boyd, Andrew D. 2025. "Towards Conversational Assistants for Health Applications: Using Chatgpt to Generate Conversations About Heart Failure."

Te Tāhū Hauora Health Quality & Safety Commission. 2024. "Co-Designing with Consumers, Whānau and Communities." Retrieved 22/Aug, 2025, from https://www.hqsc.govt.nz/consumer-hub/engaging-consumers-and-whanau/implementing-the-code/co-designing-with-consumers-whanau-and-communities/

Templin, T., Fort, S., Padmanabham, P., Seshadri, P., Rimal, R., Oliva, J., Hassmiller Lich, K., Sylvia, S., and Sinnott-Armstrong, N. 2025. "Framework for Bias Evaluation in Large Language Models in Healthcare Settings," *npj Digital Medicine* (8:1), p. 414.

University of Auckland, f. o. m. a. h. s. n.d. "Health Data Platform." from https://www.healthdata.auckland.ac.nz/

Van Citters, A. 2017. "Experience-Based Co-Design of Health Care Services," Institute for Healthcare Improvement.

Vimbi, V., Shaffi, N., and Mahmud, M. 2024. "Interpreting Artificial Intelligence Models: A Systematic Review on the Application of Lime and Shap in Alzheimer's Disease Detection," *Brain Inform* (11:1), p. 10.

Yale School of Medicine. 2024. "Bias in, Bias Out: Tackling Bias in Medical Artificial Intelligence." Retrieved 20/Aug, 2025, from https://medicine.yale.edu/news-article/bias-in-bias-out-yale-researchers-pose-solutions-for-biased-medical-ai/

Yogarajan, V., Dobbie, G., Leitch, S., Keegan, T. T., Bensemann, J., Witbrock, M., Asrani, V., and Reith, D. 2022. "Data and Model Bias in Artificial Intelligence for Healthcare Applications in New Zealand," *Frontiers in Computer Science* (Volume 4 - 2022).

Acknowledgements

The authors would like to thank Sun Eui Lee for her valuable assistance in organising and preparing the research materials that supported this study. Her contribution was essential in facilitating the smooth progression of the project.

Copyright