INFORMATION FUSION AND LEARNING FROM DIVERSE CLINICAL COHORTS AND HETEROGENEOUS DATA

Pablo-Miki Martí Castellote

TESI DOCTORAL UPF / 2023

Thesis supervisors

Dr.Bart Bijnens, Dra. Gemma Piella,

DEPARTMENT OF COMMUNICATION AND INFORMATION TECHNOLOGIES

upf. Universitat Pompeu Fabra Barcelona

To my family and Carmen

In appreciation

These past four years have been a steep learning curve for me. Not only have I learned to deal with complex research questions and to face dead ends, but I've also grown to appreciate the beauty of science.

A big thanks to Bart and Gemma for their guidance. They offered me exciting opportunities and kept me on the path to the finish line when I veered off course.

Sergio deserves a special mention for being a role model, his kindness, patience, and high-quality standards, not to mention his positive energy have made me a better researcher. In the same sense, I wanted to acknowledge Filip, Mariana, and Guillermo.

To all my colleagues at DTIC, Becaris Precaris, and Camareros: you've made this journey much more enjoyable with your company.

I wanted to thank Devyani for including me in projects with real-world impact, which are a source of motivation.

To all the folks in Boston, particularly Scott, Brian, and Jon, I'm excited to join you. I’m extremely thankful to Scott for letting me conduct research in a place so full of talented people and exciting projects. I also wanted to thank Brian for his clear insights and for making statistics so engaging. And lastly, I wanted to thank Jon for his support and shared excitement about the potential of AI we’ll start to explore soon.

These are but brief acknowledgements of the many who have contributed to my PhD journey. I'm grateful for all the shared wisdom and look forward to continuing exploring.

Abstract

This PhD thesis explores how data science can be leveraged to learn from diverse cohorts and fuse information from heterogeneous data, aiming to advance understanding and improve prediction in healthcare settings. It covers a broad set of scenarios, namely: It evaluates Machine Learning's capabilities in enhancing disease phenotyping and clinical trial efficiency, the prediction of adverse perinatal outcomes in cross-cultural cohorts, the understanding of the impact of interventions on patient-centered outcomes, and also provides a technical contribution through a novel incremental Multiple Kernel Learning technique. The research illustrates the potential of data science when applied to complex cohorts and data, thereby offering innovative methodologies and insights for managing large clinical datasets.

Keywords: Information Fusion, Heterogeneous Data, Data Science, Healthcare

Resumen

Esta tesis de doctoral explora cómo aprovechar la ciencia de datos para aprender de cohortes diversas y fusionar información de datos heterogéneos, con el objetivo de avanzar en la comprensión y mejorar las predicciones en entornos de atención médica. Cubre un amplio conjunto de escenarios, a saber: Evalúa las capacidades del aprendizaje automático para mejorar la fenotipado de enfermedades y la eficiencia de los ensayos clínicos, la predicción de eventos perinatales adversos en cohortes interculturales, la comprensión del impacto de intervenciones terapéuticas en el contexto vital del paciente, y también proporciona una contribución técnica a través de una nueva técnica de Aprendizaje de Múltiples Kernels de forma incremental. La investigación ilustra el potencial de la ciencia de datos cuando se aplica a cohortes y datos complejos, ofreciendo así metodologías e ideas innovadoras para gestionar grandes conjuntos de datos clínicos.

Palabras clave: Fusión de Información, Datos Heterogéneos, Ciencia de Datos, Sistemas de Salud.

Introduction

Context and motivation

In this thesis we focus on open problems in areas of the healthcare system that could be very positively impacted by technological innovations based on information fusion and machine learning (ML).

Healthcare systems worldwide are grappling with severe pressures, a struggle that is particularly pronounced in low and lower-middle-income countries (1). These pressures predominantly arise from workforce deficiencies and budget constraints. Nonetheless, the vast amount of patient data that healthcare systems generate and collect during their functioning has been shown to be an underutilized resource with the potential to alleviate these pressures. These data, given the proper processing, could be leveraged to optimize and streamline many processes in the system. For instance, it may help clinicians in the decision-making process of diagnoses and treatment plans, ultimately standardizing care (2). However, the challenge lies in the integration of multiple data modalities such as imaging-derived features, clinical test biomarkers, and other descriptors of a patient's status, which is not a straightforward process (3). Therefore, in this thesis, we aim to address this issue by presenting practical cases that apply information fusion and machine learning (ML) techniques based on patient data to improve healthcare outcomes.

As a brief primer, information fusion in the context of this thesis refers to the process of integrating multiple data sources to produce information that is more comprehensive and useful than that provided by any individual data source. In our scenarios, this relates to confronting distinct clinical cohorts, harmonizing multimodal patient data, or incorporating information from different psychosocial aspects of a patient's life. Moreover, ML refers to a broad set of potent algorithms and statistical models that extract patterns from vast amounts of data and may perform inferences on new cases.

As per the clinical part, we focus on the fields of cardiovascular medicine, and fetal and pediatric medicine. The rationale for working in these fields is twofold. Firstly, cardiovascular medicine holds global significance as heart disease currently stands as the leading cause of mortality worldwide. The aging population and escalating lifestyle-related risk factors, such as obesity and diabetes, accentuate its importance (4). Secondly, fetal and pediatric medicine provide an opportunity for early-life interventions, which could potentially prevent lifelong health complications and enhance quality of life (5).

Within the defined fields, the specific areas of our contributions encompass the evidence generation process for therapies, preventive medicine, and value-based healthcare. We will now illustrate the motivation for our contributions in the three of them.

First, the gold standard for evidence generation for therapies is randomized clinical trials (RCT). During an RCT, two groups are recruited where one of them receives either placebo or the current standard of care, and the other receives the therapy to be tested. The group sizes need to be sufficiently large to allow for statistical significance. This design causes trials to struggle with several aspects. For instance, patient recruitment requires stringent inclusion criteria, making the identification of appropriate candidates a laborious task, and associated risks to participants make participation unappealing. This is further compounded by high operational costs attributed to trained staff and data collection infrastructure. These factors limit the recruitment of large cohorts, thus making it hard to construct bespoke therapies for each patient phenotype. Instead, expert panels are compelled to generalize the observed heterogeneity into broader groups. This provides physicians with practical solutions, albeit potentially less individually tailored (6). General population data, if properly processed and understood, may hold the key to reduce the cost of trials by getting recycled, leading to a significantly smaller recruitment requirement (7).

Regarding the second area, preventive medicine plays a crucial role in reducing avoidable deaths. This is particularly crucial when preventing adverse perinatal events. To that end, the identification of high-risk pregnancies can enable timely interventions to prevent fatal and avoidable outcomes for mothers and children. As previously reported by the World Health Organization (WHO), perinatal adverse events remain a critical yet elusive challenge in both high- and low-income settings, where the latter further struggles with compounding factors such as undernutrition and limited access to healthcare (8). Integrating cohorts from both high- and low-income countries together with comprehensive descriptors from patients might give insights into the etiologies and help identify actionable items to improve care.

Lastly, regarding value-based healthcare – a care delivery model that shifts the focus from the traditional 'fee-for-service' model (where providers are paid based on the amount of healthcare services they deliver) to a 'fee-for-value' model (where providers are compensated based on the quality of patient outcomes and cost-effectiveness) – integrating a multitude of patient perspectives is vital (9). A process that involves collecting extensive questionnaires about patients' psychosocial statuses can assist in quantifying the social impact of treatments and identify areas that either show pronounced improvement or face challenges. This holistic evaluation goes beyond traditional healthcare metrics, as it includes factors like social integration, mental wellbeing, and overall quality of life. By taking such a comprehensive approach, we can highlight the intricate relationship between healthcare delivery and the societal context in which it exists.

Objectives and proposed approaches

This thesis employs a diverse set of data science techniques—classical statistical methods, supervised learning, and unsupervised learning—to solve the problems shown in the previous section. We define four primary objectives and briefly explain the implemented approach to attain it.

Objective 1: Improve Clinical Trial Efficiency using General Population Data.

We focus on the case of heart failure, where we sift through general population data with the aim of identifying individuals who exhibit cardiac function similar to that of trial participants. These individuals can act as a synthetic control arm, thereby reducing costs, workload, and patient risks in clinical trials. We critically choose Unsupervised Multiple Kernel Learning (uMKL), an unsupervised machine learning algorithm, to comprehend the phenotypic space of heart failure and perform comprehensive pairwise matching.

Objective 2: Identify Small for Gestational Age (SGA) using ML in Different Cultural Cohorts.

We aim at classifying pregnancies into SGA or controls at a point in the pregnancy with no clear clinical assessment. To do so, we train extreme gradient boosting trees from maternal clinical data, fetal biometry, and blood flow measurements. This approach allows us to examine the predictive role of these descriptors across distinct populations, evaluating the model's transferability and generalizability. This comparison provides insights into population-specific etiologies, indicating areas where model adjustments may be necessary.

Objective 3: Assess the Social Impact of Clubfoot Treatment in India.

We aim at obtaining a holistic understanding of families affected by clubfoot and devising approaches to measure the impact of treatment interventions. To do so, we make use of classical statistical methods, which validate the findings derived from questionnaires. This will allow us to identify challenges faced by families of children who have undergone clubfoot treatment. The outcomes will guide future research directions, including a long-term follow-up study to assess the broader impacts of the treatment beyond health outcomes.

Objective 4: Enhance the scalability of Unsupervised Multiple Kernel Learning.

To overcome dataset size limitations in uMKL, we propose the development of an incremental learning extension. This modification allows the model to integrate new observations without the need for batch mode retraining.

The previous objectives deploy one or more of the three primary data science techniques. Supervised ML, used in Objective 2, facilitates the creation of decision support systems capable of detecting subclinical changes for early intervention. Unsupervised ML, utilized in Objectives 1 and 4, enables the model to independently explore patient data, uncovering hidden relationships and disease presentations. Classical statistical methods, fundamental to Objective 3, help quantify interpretable associations between variables and test for population differences.

Thesis outline

The research conducted within this thesis is structured in self-contained chapters formatted as research journal papers, and we list the remaining chapters below.

Chapter 1: Exploring registry and trial data with Machine Learning.
In this chapter we propose an ML framework that can play a role in streamlining Heart Failure clinical trials, a syndrome known for its heterogeneity, to enhance their efficiency.

Chapter 2: Cross-cultural Machine Learning for predicting adverse perinatal outcomes.
In this chapter we consider two distinct populations of pregnant women, one from Barcelona and one from Pakistan, to predict factors that can contribute to adverse perinatal events on time.

Chapter 3: Assessing the Impact of Interventions of Club Foot India Initiative Trust on Patients and Their Families.
This chapter will focus on the practical application of social and data science to assess the impact of the Ponseti treatment in clubfoot children provided by the non-governmental organization CURE India Initiative Trust.

Chapter 4: Incremental Multiple Kernel Learning
This final chapter provides a technical contribution that will address the lack of scalability of batch unsupervised Multiple Kernel Learning (uMKL).

Conclusion: We present a summary of contributions from this thesis alongside limitations and future research directions.

Table of contents

List of figures

List of supplementary figures

List of tables

List of supplementary tables

1. EXPLORING REGISTRY AND TRIAL DATA WITH MACHINE LEARNING

Abstract

a) Background

Clinical trials are burdened by the need for large cohort numbers to demonstrate treatment effect and large and complex datasets commonly difficult to integrate and analyze. Machine learning (ML) can be used to unravel the complexity and improve handling of clinical trial data. We aim to demonstrate the capacity of ML in combining heterogeneous patient data to improve cohort phenotyping, fusing datasets to reinforce findings, and increasing the efficiency of running trials by '*recycling'* existing data through advanced patient matching of patients from the general population to those in trials.

b) Methods and results

Subjects from the general population were included from the Atherosclerosis Risk in Communities (ARIC) study (n=2123), as well as patients presenting with heart failure with reduced ejection fraction from the Multicenter Automatic Defibrillator Implantation Trial - Cardiac Resynchronization Therapy (MADIT-CRT) (n=429), and those with preserved ejection fraction from the Treatment of Preserved Cardiac Function Heart Failure with an Aldosterone Antagonist (TOPCAT) trial (n=218). The main inclusion criteria for our analysis were the availability of echo data and 4- and 2- chamber

speckle-tracking analysis. Outcome was defined as death from any cause or heart failure hospitalization, whichever came first. ML input consisted of 12 segmental left ventricular (LV) strain curves, an LV volume curve, and a set of 16 demographic, medical history, and clinical parameters. The first step of the ML analysis consisted of obtaining the ML-derived space for the ARIC cohort, defining the spectrum of disease in the general population, since patients with similar data will be positioned close to each other in the ML-space. In the second step, TOPCAT and MADIT-CRT patients were 'projected' into the ML-derived space using the identical input features, and clustering was performed to define phenogroups. Finally, a "synthetic control arm" was obtained by matching subjects from the general population with trial patients based on location in the ML-derived space. These selected ARIC patients were then compared with the trial ones based on clinical characteristics, as well as cardiac mechanics.

The ML algorithm positioned ARIC subjects into a common space based on integrated clinical and whole cardiac cycle echo data. Feature maps of the ARIC ML-derived space showed subject separation based on physiologically sensible patterns, and in relation to differing clinical risk of outcome. The addition of heart failure patients confirmed the spectrum of disease defined in the general population and highlighted the high-risk regions, whereas clustering defined three clinically distinct patient phenotypes. Patient matching using unsupervised learning successfully produced patient pairs with similar cardiac mechanics and clinical characteristics, better than when compared to patients selected using history of heart failure and ejection fraction.

c) Conclusion

The presented analysis serves as a proof-of-concept for an unsupervised ML approach in analyzing registry and trial datasets. ML can enhance the integration of patient data and improve disease phenotyping, fuse different patient cohorts to build trust towards findings, and potentially increase the efficiency of running clinical trials by advanced patient matching through the ‘recycling’ of existing patient datasets.

1.1 Introduction

Current clinical trials offer the gold standard to measure the impact of healthcare interventions on patient populations. Generally, trials focus on a specific disease, where the aim is to collect detailed clinical information and integrate it to recognize patient phenotypes. A certain phenotype, selected based on specific inclusion/exclusion criteria, is then randomized into treatment and control arms, with as main goal the generation of novel medical insights (Figure 1**,** left). The clinical trial process is burdened by the need for large cohort numbers to demonstrate treatment effect and large and complex datasets that are commonly difficult to integrate and analyze, resulting in high running costs as well as inefficiencies in recruitment and data usage (10). It is also necessary to keep in mind that data monitoring in clinical trials is essential for meeting quality and ethical standards, as well as regulatory compliance. However, it can be a complex and costly process that poses a burden on trial staff and participants, in addition to associated risks for the latter. The burden on trial participants can include additional procedures, stress, and inconvenience (11,12**).** Machine learning can help streamline many aspects of clinical trials (Figure 1, right)(13). Indeed, stakeholders such as the US Federal Drug Agency have been pushing towards the modernization of clinical trials, identifying an important role for computer modelling and in-silico methods - from models predicting product safety and efficiency, virtual physiological patients testing medical products, trial simulations revealing patient-therapy and disease interactions, to knowledge building tools aiding data mining(14,15).

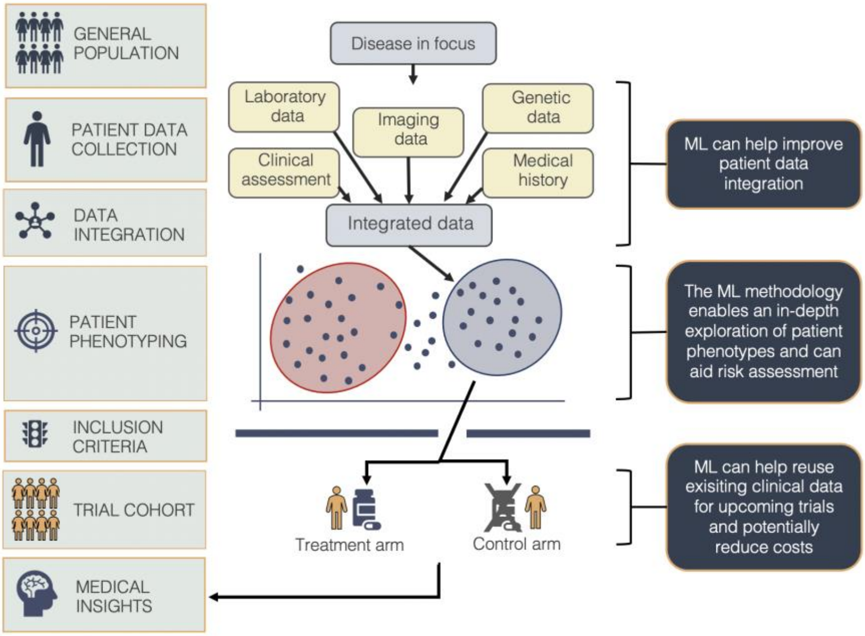


Figure 1 Clinical trial workflow and the potential targets of ML

We hypothesize that unsupervised ML can be used as a tool to unravel the complexity as well as improve handling of clinical trial data. We aim to demonstrate the capacity of manifold learning (implemented through multiple kernel learning (MKL)) in integrating heterogeneous patient data from a general population registry cohort - combining demographic, comorbidity, imaging, and laboratory information- to improve patient phenotyping and enable direct patient comparisons(16,17). Furthermore, we will showcase the utility of ML in fusing registry and trial data to reinforce findings learned through data integration within the general population. Finally, similarly to propensity-score matching, we will demonstrate how ML could increase the efficiency of running trials by '*recycling'* existing data through advanced patient matching of patients from general population cohorts to those in trials in order to create a synthetic control arm.

1.2 Materials and methods

A general overview of the methodology is shown in Figure 2. A detailed description of the ML methodology (data selection, preprocessing steps, MKL, and the analysis of the ML-derived space) can be found in the Supplementary Materials.

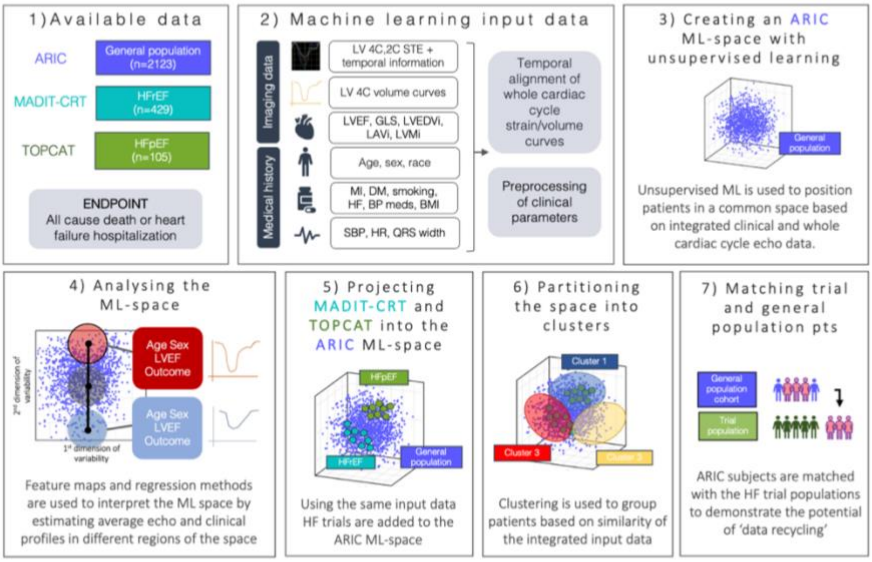

1) Available data
ARIC
General population (n=2123)
MADIT-CRT
HFrEF (n=429)
TOPCAT
HFpEF (n=105)
ENDPOINT
All cause death or heart failure hospitalization
2) Machine learning input data
Imaging data
Medical history
LV 4C,2C STE + temporal information
LV 4C volume curves
LVEF, GLS, LVEDVi, LAVi, LVMi
Age, sex, race
MI, DM, smoking, HF, BP meds, BMI
SBP, HR, QRS width
Temporal alignment of whole cardiac cycle strain/volume curves
Preprocessing of clinical parameters
3) Creating an ARIC ML-space with unsupervised learning
General population
Unsupervised ML is used to position patients in a common space based on integrated clinical and whole cardiac cycle echo data.
4) Analysing the ML-space
2nd dimension of variability
1st dimension of variability
Age Sex LVEF Outcome
Age Sex LVEF Outcome
Feature maps and regression methods are used to interpret the ML space by estimating average echo and clinical profiles in different regions of the space
5) Projecting MADIT-CRT and TOPCAT into the ARIC ML-space
HFpEF
HFrEF
General population
Using the same input data HF trials are added to the ARIC ML-space
6) Partitioning the space into clusters
Cluster 1
Cluster 3
Cluster 2
Clustering is used to group patients based on similarity of the integrated input data
7) Matching trial and general population pts
General population cohort
Trial population
ARIC subjects are matched with the HF trial populations to demonstrate the potential of 'data recycling'

Figure 2 Overview of methodology

4CH- 4-chamber; 2CH – 2-chamber; LV left ventricle, LVEF – LV ejection fraction; GLS – global longitudinal strain; LVEDVi - LV end-diastolic volume indexed to body surface area; LAVi - left atrial volume indexed to body surface area, LVMi – LV mass indexed to body surface area; MI – myocardial infarction; DM – diabetes mellitus; HF – heart failure; BP – blood pressure; BMI – body mass index; SBP – systolic blood pressure; HR – heart rate

a) Patient Cohorts

Three independent cohorts were included in the analysis. Patients from the general population were included from the Atherosclerosis Risk in Communities (ARIC) study (18,19), those presenting with heart failure with reduced ejection fraction (HFrEF) from the Multicenter Automatic Defibrillator Implantation Trial -

Cardiac Resynchronization Therapy (MADIT-CRT) trial (20,21), whereas, patients presenting with heart failure with preserved ejection fraction (HFpEF) from the Treatment of Preserved Cardiac Function Heart Failure with an Aldosterone Antagonist (TOPCAT) trial (22). The main inclusion criteria for our analysis were the availability of echo data and the 4- and 2- chamber speckle-tracking segmentations. ARIC data was obtained at visit 5, whereas data from TOPCAT and MADIT was obtained at randomization, i.e. prior to providing treatment to the treatment arm, thus we assume no difference between patients by treatment assignment. This resulted in 4246 participants from the ARIC cohort with echocardiographic data available, from which we randomly selected half of them, resulting in 2123 participants included (35% of the cohort, average follow-up 2.6 years). This selection was done to reduce computational costs. From MADIT-CRT, 429 patients were included (24% of the cohort, average follow-up 2 years), 218 patients from the TOPCAT trial (6% of the cohort, average follow-up 2.9 years). Outcome was defined as death from any cause or heart failure (HF) hospitalization, whichever came first, with the adjudication of endpoints previously described (19,20,23) (Figure 2, box 1).

b) Echocardiographic Data

Demographic data, medical history data, as well as blood pressure and ECG measurements were available in all cohorts. Echocardiographic data analysis was performed at the Echocardiography Reading Center (*ERC; Brigham and Women's Hospital*, *Boston*, *MA*), as according to guidelines (24). Speckle tracking (STE) deformation analysis was performed using the TomTec Arena software (*v1.0, TomTec Imaging Systems, Unterschleissheim, Germany*), with the endocardial border traced in the end-systolic frame of the apical 4- and 2-chamber views (4CH and 2 CH), and the segmental left ventricular (LV) longitudinal strains and a volume curve exported as text files. The analysis was approved by the institutional review board at each of the participating centers.

Data available in all three cohorts was used as the ML algorithm input. This included the 4CH and 2CH strain curves (12 LV segments), LV volume curves, and the available demographic, medical history, and clinical parameters (age, sex, race, history of myocardial infarction, diabetes, smoking, history of heart failure, blood pressure medications, body mass index (BMI), heart rate, systolic blood pressure (SBP), QRS width on the ECG, 4CH global longitudinal strain (4CH-GLS), LV end-diastolic volume indexed to body surface area (BSA) (LVEDVi), left atrial volume indexed to BSA (LAVi), and LV mass indexed to BSA (LVMi)). To allow for inter-subject comparisons due to inter-patient variability in heart rate, the strain curves were temporally aligned using diffeomorphic registration (25,26). Volume profiles were centered at zero and registered to a reference to match the cardiac cycle phase. Missing clinical parameters were imputed by using factor analysis of mixed data (imputeFAMD function in the missMDA package in R 4.2.1) (Figure 2, box 2). In summary, the combination of strain/volume curves and the clinical parameters comprised a total of 1485 data points for every subject.

c) Dimensionality Reduction and Clustering

Unsupervised MKL (27)- implemented using MATLAB (R2022a, 2022, The MathWorks Inc., Natick, MA, USA) was then used to integrate the 1485 data points per subject and position them in a virtual space representation of their integrated clinical and echo characteristics. This concept is called dimensionality reduction and involves integrating and transforming heterogeneous patient data into a location within a common, low-dimensional representative space (referred to as 'ML-derived space' in the continuation of the text). The axes of this ML-derived space (denoted as Dimensions 1, 2, 3, etc.) represent the distinct dimensions of the patient data - each dimension captures a specific aspect of the input data characteristics (for example, a particular dimension of the ML-derived space may distinguish patients based on factors such as age, gender, or reduced 4CH-GLS). Therefore, patients are positioned based on their integrated characteristics, but blinded in relation to their original cohort or heart failure status. Patients with similar echo and clinical characteristics will be positioned closer together, whereas those with opposing characteristics further apart.

The first step of our methodology consisted of obtaining the ML-derived space on the ARIC population, thus defining the spectrum of normalcy and disease in the general population (Figure 2, box 3). The analysis of this ML-derived space was performed through feature maps and regression analysis (Figure 2, box 4). Feature maps imply color-coding the ML-derived space for specific patient characteristics (e.g., age, QRS width, 4CH-GLS) and creating endpoint heatmaps (kernel density maps used to define high risk regions), whereas regression analysis (multiscale kernel regression) enabled the estimation of average deformation and volume profiles in different points within the space (points were placed geodetically, i.e., following the shape of the space).

Once the general population characteristics were explored, the TOPCAT and MADIT-CRT patients were 'projected' into the ML-derived space using identical input features (Figure 2, box 5). 'Projecting' implies the addition of new subjects to a ML-derived space already created with ARIC data without retraining with the new information. The ML-derived space was now once again explored (through feature maps and regression analysis), and additionally, K-means clustering was performed (28) to group subjects into phenogroups with similar characteristics. The number of clusters was predetermined to be 3 (the number of expected clinical categories), and the average volume and deformation profiles, clinical characteristics, and cluster-wise event rates were reported for each of the resulting clusters (Figure 2, box 6).

d) Synthetic Controls

Finally, the last step of the analysis matched subjects from the general population to trial patients, creating a potential synthetic control arm (Figure 2, box 7). To achieve this, we compared two matching approaches for each TOPCAT and MADIT patient: ML matching and threshold-based matching. On the one hand, ML matching involved identifying the nearest ARIC subject in the ML-derived space. On the other hand, threshold-based matching selected the closest ARIC patient based on history of HF and ejection fraction

(EF). Both approaches were implemented using sampling without replacement – meaning that a single ARIC patient will only be matched to a single trial patient.

d) Statistical Analysis

When assessing the differences in variables between groups we proceed as follows: binary variables were analyzed by calculating relative frequencies within each group, followed by Chi-square tests to evaluate statistical significance. Continuous variables were subjected to a Kolmogorov-Smirnov test to determine their distribution. For normally distributed variables, the mean, standard deviation, and an ANOVA test were performed. Conversely, non-normally distributed variables were analyzed using median, first and third quartiles, and a Kruskal-Wallis test.

1.3 Results

The baseline clinical and echo characteristics, as well as the outcomes are shown in Table 1.

Table 1 Clinical and echo characteristics of the cohorts

	ARIC (n=2123)	**TOPCAT (n=218)**	**MADIT-CRT (n=429)**
Outcome, n (%)	122 (5.75%)	29 (27.62%)	109 (25.41%)
Age, years (IQR)	74.88 (71.5 to 79.4)	71.19 (62.29 to 78.26)	65 (57 to 71.25)
Female sex, n (%)	1221 (57.51%)	115 (52.75%)	104 (24.24%)
Caucasian, n (%)	1751 (82.48%)	169 (77.52%)	391 (91.14%)
BMI, kg/m2 (IQR)	27.23 (24.51 to 30.48)	30.37 (26.57 to 34.93)	28 (25.08 to 30.7)
Diabetes, n (%)	601 (28.31%)	64 (29.36%)	120 (27.97%)

Current smoker, n (%)	132 (6.22%)	27 (12.39%)	56 (13.05%)
Myocardial infarction, n (%)	23 (1.08%)	56 (25.69%)	182 (42.42%)
Prior heart failure, n (%)	21 (0.99%)	132 (60.55%)	157 (36.6%)
Blood pressure medication, n (%)	1508 (71.03%)	166 (76.15%)	411 (95.8%)
Systolic blood pressure, mmHg (IQR)	129 (119 to 140)	126.71 ± 15.23	120 (108 to 132)
Heart rate, n (%)	61 (55 to 67)	66 (60 to 74)	63 (56 to 70)
QRS width, ms (IQR)	90 (84 to 100)	92.5 (84 to 108)	158 (142 to 170)
LVMi, g/m2 (IQR)	74.61 (65.07 to 87.28)	99.36 (82.59 to 119.36)	104.63 (93.05 to 116.28)
LVEDVi, ml/m2 (IQR)	42.42 (36.69 to 49.4)	47.19 (37.89 to 55.18)	120.01 (108.01 to 137.91)
LAVi, ml/m2 (IQR)	24.39 (20.15 to 29.33)	29.04 (21.7 to 35.09)	45.5 (39.76 to 52.62)
LVEF, % (IQR)	66 (62.3 to 69.4)	60 (55 to 65)	25 (20 to 28)
4CH-GLS, % (IQR)	-18.87 (-20.27 to -17.33)	-17.18 ± 3.11	-9.45 (-11.08 to - 7.78)

IQR – interquartile range; BMI – body mass index; LVMi – LV mass indexed to body surface area; LVEDVi - LV end-diastolic volume indexed to body surface area; LAVi - left atrial volume indexed to body surface area, LVEF – LV ejection fraction; 4CH- 4-chamber; GLS – global longitudinal strain.

MKL positioned ARIC subjects into a common space based on integrated clinical and whole cardiac cycle echo data (Figure 3A). Feature maps showed Dim 1 separated patients based on sex, BMI, heart rate and LVEDVi and LAVi, whereas Dim 2 separated patients based on a gradient of systolic impairment (4CH-GLS and EF) (Figure 3B). Regression demonstrated distinct cardiac mechanic patterns defined within Dim 2 – ranging from normal 4CH-GLS values (Figure 3C, orange), a pattern of intraventricular dyssynchrony in the form of a septal flash and lateral basal wall stretching (Figure 3C, green), and overall reduced 4CH-GLS (Figure 3C, red). Endpoint heatmaps defined the lower region of Dim 2 (in the continuation of the text referred to as the ‘ARIC tail’) to be at higher risk of primary outcome (Figure 3D).

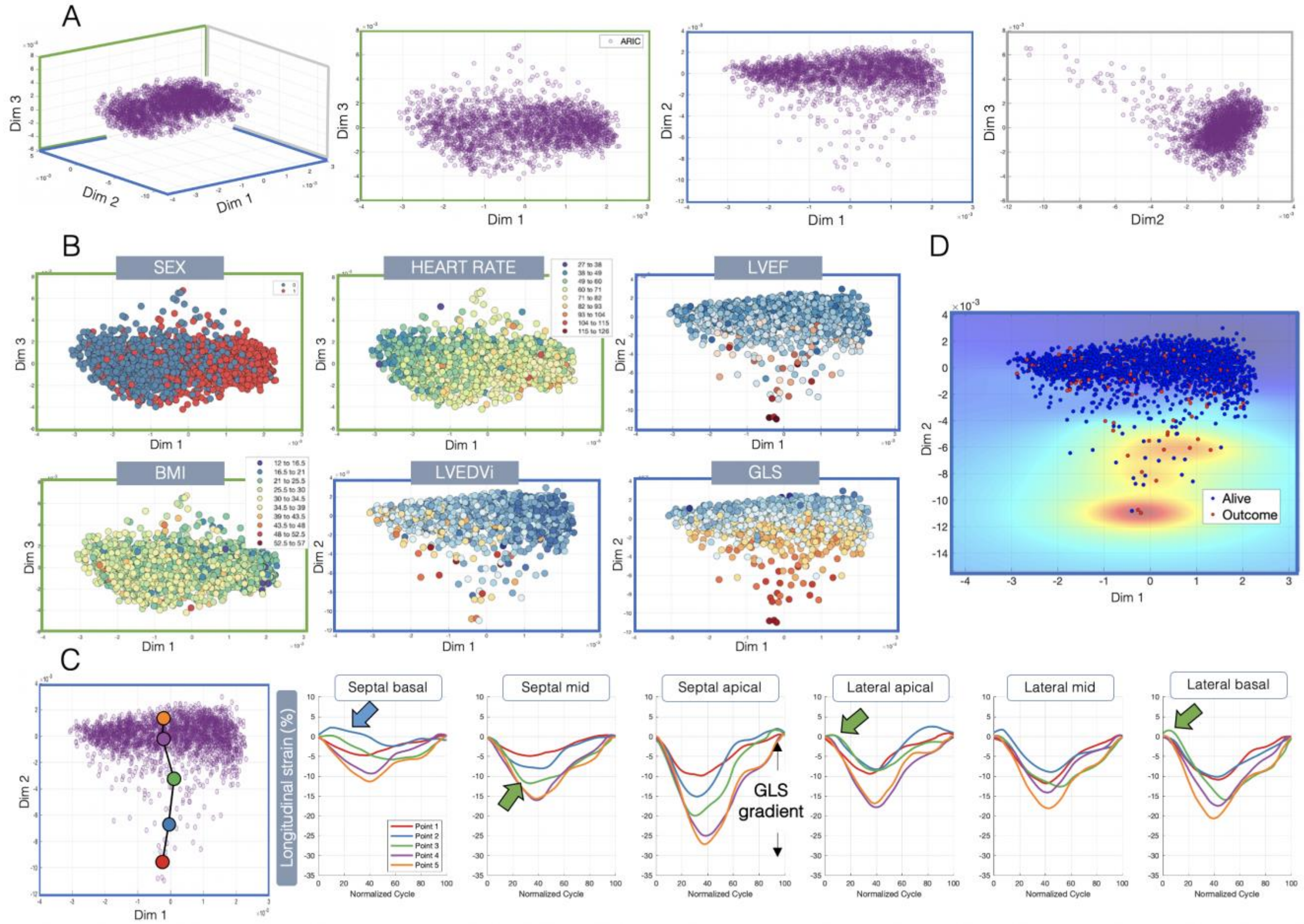
A
Dim 3
Dim 2
Dim 1
ARIC
Dim 3
Dim 1
Dim 2
Dim 1
Dim 3
Dim2
B
SEX
0
1
Dim 3
Dim 1
HEART RATE
27 to 38
38 to 49
49 to 60
60 to 71
71 to 82
82 to 93
93 to 104
104 to 115
115 to 126
Dim 3
Dim 1
LVEF
Dim 2
Dim 1
BMI
12 to 16.5
16.5 to 21
21 to 25.5
25.5 to 30
30 to 34.5
34.5 to 39
39 to 43.5
43.5 to 48
48 to 52.5
52.5 to 57
Dim 3
Dim 1
LVEDVi
Dim 2
Dim 1
GLS
Dim 2
Dim 1
D
×10⁻³
Dim 2
Dim 1
Alive
Outcome
C
Dim 2
Dim 1
Longitudinal strain (%)
Septal basal
Point 1
Point 2
Point 3
Point 4
Point 5
Normalized Cycle
Septal mid
Normalized Cycle
Septal apical
GLS gradient
Normalized Cycle
Lateral apical
Normalized Cycle
Lateral mid
Normalized Cycle
Lateral basal
Normalized Cycle

Figure 3 Analysis of the ARIC based ML-derived space

A) Three-dimensional representation of the ARIC -based ML-derived space. The axes represent the first three dimensions of variability of the integrated clinical and whole-cardiac cycle data (i.e., each dimension of the space separates patients based on certain differences in input data). Patients with similar characteristics are positioned closer together, whereas those with different characteristics further apart. Two-dimensional views of the same ML space are shown on the right, with color-coded frames to relate to the three-dimensional space (e.g., the two-dimensional view of the ARIC population shown in the blue frame is the arial view of the three-dimensional space). B) Feature maps of the space are shown together with color coded frames that help relate to the 3-dimensional space in A). The feature maps show the separation of patients based on characteristics across the first three dimensions. C) By using regression methods, we can identify and directly compare the average deformation profiles in different parts of the ML-derived space. Points are positioned across Dimension 2 geodesically, i.e., following the shape of the patient population. The average deformation curves of the 6 LV segments in the 4-chamber view are shown on the right. The x-axis represents the normalized cardiac cycle from aortic valve opening to mitral valve closing (0-100%), and the y-axis represents the % for strain. These deformation curves are color-coded corresponding to the same-colored points in the ML-derived space (i.e., red curves describe the average strain of the region marked with the red point). The analysis captures different cardiac mechanic patterns within the population and explains the basis upon which Dimension 2 separates patients. The blue arrow marks a regional deformation abnormality in the region marked by the blue point, potentially related to ischemic disease. The green arrows show a pattern of septal flash and stretching of the lateral wall in early systole, and contraction of the lateral wall and stretching of the septal wall in late systole. Finally, the peak longitudinal strain values show a gradient across Dimension 2, as is best seen in the septal apical segment, confirming the finding of Dimension 2 encoding strain gradient as visualized in B). D) An outcome heat map shows the lower parts of Dimension 2 ('ARIC tail') as a region with a higher percentage of outcomes. A more detailed explanation of the methodology can be found in the ***Supplement****.*

After defining the spectrum of changes within the ARIC population, HFrEF patients were added to the ML-derived space, projecting in the mid region of the first three dimensions, whereas the HFpEF patients projected in between the ARIC and MADIT-CRT cohorts (Figure 4A). With the addition of HF patients, the new feature maps upheld the distribution of patient characteristics, accentuated gradients of LV enlargement and dysfunction, defined a clear region of QRS prolongation in the central region, and highlighted the lower dimension of Dim2 (i.e., 'ARIC tail') as a high-risk region (Figure 4B and C).

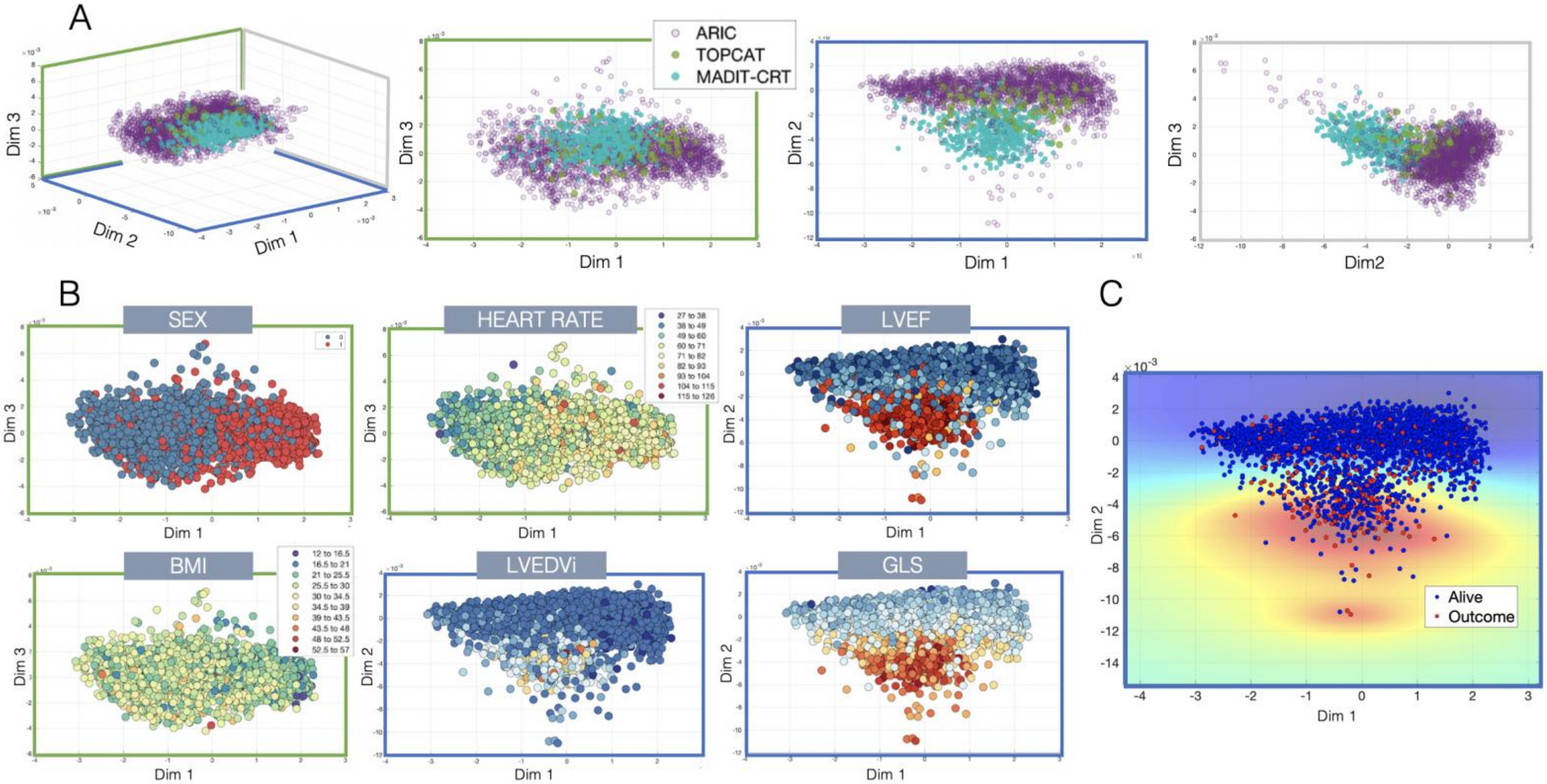


Figure 4 Analysis of the fused cohort ML-derived space

A) ARIC-based ML-derived space after projection of the TOPCAT and MADIT-CRT cohorts. The 2-dimensional views visualize the separation of the three patient groups, although, expectedly, with a considerable overlap in bordering regions. B) The feature maps of the fused cohorts validate the findings in Figure 3B. C) The outcome heatmap shows a better-defined high-risk region after the addition of 'high-outcome' HF patients overlapping with the one seen in Figure 3D.

K-means clustering defined three patient phenotypes within the combined cohort (Figure 5, Table 2).

Table 2 Clinical and echo characteristics of the three ML-derived clusters

	Cluster 1 (n=1355)	**Cluster 2 (n=859)**	**Cluster 3 (n=443)**	**P-value**
ARIC, n (%)	1297 (95.72%)	736 (85.68%)	90 (20.32%)	<0.001
TOPCAT, n (%)	46 (3.39%)	40 (4.66%)	19 (4.29%)	0.307
MADIT, n (%)	12 (0.89%)	83 (9.66%)	334 (75.4%)	<0.001
Outcome, n (%)	72 (5.31%)	70 (8.15%)	118 (26.64%)	<0.001
Age, years	75.32 ± 5.38	74.61 ± 7.05	66.54 ± 11.53	<0.001
Female sex, n (%)	808 (59.63%)	456 (53.08%)	123 (27.77%)	<0.001
Caucasian, n (%)	1125 (83.03%)	712 (82.89%)	391 (88.26%)	0.022
BMI, kg/m2	27.65 ± 5.01	28.37 ± 5.17	28.36 ± 5.12	<0.001
Diabetes, n (%)	350 (25.83%)	271 (31.55%)	129 (29.12%)	0.013
Current smoker, n (%)	87 (6.42%)	62 (7.22%)	50 (11.29%)	0.003
Myocardial infarction, n (%)	29 (2.14%)	45 (5.24%)	157 (35.44%)	<0.001
Prior heart failure, n (%)	41 (3.03%)	57 (6.64%)	144 (32.51%)	<0.001
Blood pressure medication, n (%)	942 (69.52%)	648 (75.44%)	410 (92.55%)	<0.001
Systolic blood pressure, mmHg	129.29 ± 17.57	130.48 ± 16.67	122.15 ± 19.15	<0.001
Heart rate, per minute	59.51 ± 8.66	65.46 ± 10.79	65.29 ± 11.53	<0.001
QRS width, ms	93.92 ± 18.08	102.97 ± 27.78	146.98 ± 32.11	<0.001
LVMi, g/m2	76.49 ± 18.02	82.93 ± 20.71	106.73 ± 23.52	<0.001
LVEDVi, ml/m2	43.97 ± 11.5	51.14 ± 24.46	110.92 ± 44.99	<0.001
LAVi, ml/m2 (IQR)	25.5 ± 7.58	26.8 ± 9.39	44.85 ± 12.13	<0.001
LVEF, %	66.36 ± 6.23	60.63 ± 13.03	31.29 ± 16.12	<0.001
4CH GLS, %	-19.68 ± 1.81	-16.94 ± 2.52	-10.05 ± 3.22	<0.001

IQR – interquartile range; BMI – body mass index; LVMi – LV mass indexed to body surface area; LVEDVi - LV end-diastolic volume indexed to body surface area; LAVi - left atrial volume indexed to body surface area, LVEF – LV ejection fraction; 4CH- 4-chamber; GLS – global longitudinal strain.

Cluster 1 (Figure 5, blue) consisted of a majority of ARIC subjects, 3% of TOPCAT patients, and less than 1% of MADIT-CRT patients. These patients were predominantly female, with non-remodeled ventricles, low BMI, and preserved LVEF and 4CH-GLS. We observed an overall low number of outcomes and a low risk profile. Cluster 2 (Figure 5, yellow) was also dominantly ARIC in population, with a similarly low part of TOPCAT, but a higher percentage of MADIT-CRT patients. The patients in this group showed a slight increase in comorbidities in comparison to Cluster 1 (e.g., higher BMI, higher percentage of prior HF and diabetes, and higher average heart rate) and a reduced 4CH-GLS despite a preserved LVEF. Cluster 3 (Figure 5, red) included a majority of MADIT-CRT patients, however with a high percentage of general population ARIC subjects. These patients were smokers, with a medical history of myocardial infarction and HF, well-regulated arterial hypertension, and remodeled LV with enlarged LA. They had overall decreased global LV function, resulting in a high risk of outcome. The ARIC outliers within this ‘ARIC tail’ had higher age and reduced longitudinal strain, potentially representing a subgroup of HF.

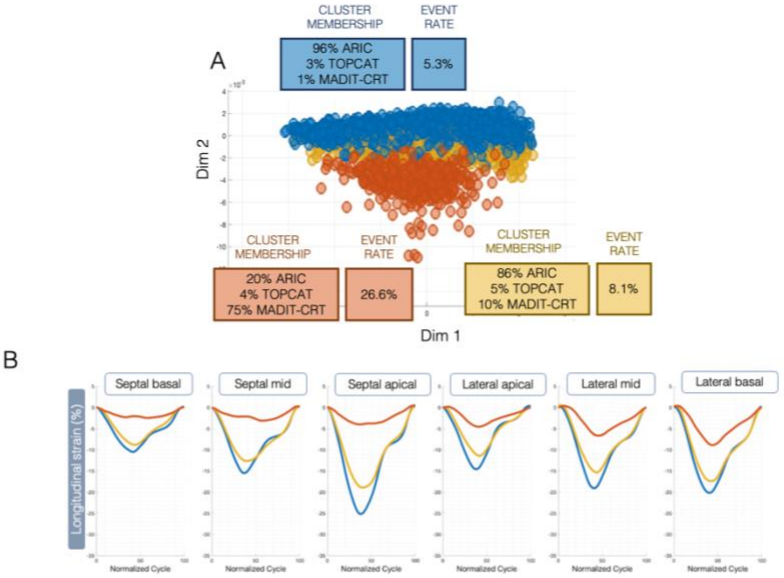


Figure 5 Clustering and further experiments within the ML-derived space

A) A 2-dimensional view of the ML-derived space shown in Figure 4 after K-means clustering. The cluster membership and event rate are shown for each of the three clusters. B) The average deformation of the 6 LV segments seen in the 4-chamber view shows clear differences in myocardial deformation within the clusters.

Finally, upon matching each HF trial patient with an ARIC counterpart, we assessed the similarity in strain, volume, and clinical and demographic variables (Figures 6 and 7 and Supplementary Tables 1 and 2). Our findings indicate that the ML matching approach yielded a selection of ARIC patients with much higher overlap in regard to cardiac mechanics and similar clinical profile than just looking at history of HF and EF. We identify ARIC patients with similar risk profiles to trial participants, yet, due to a sparse high-risk population in the ARIC subset, our matching approach results in a lower mortality rate than seen in the trial groups.

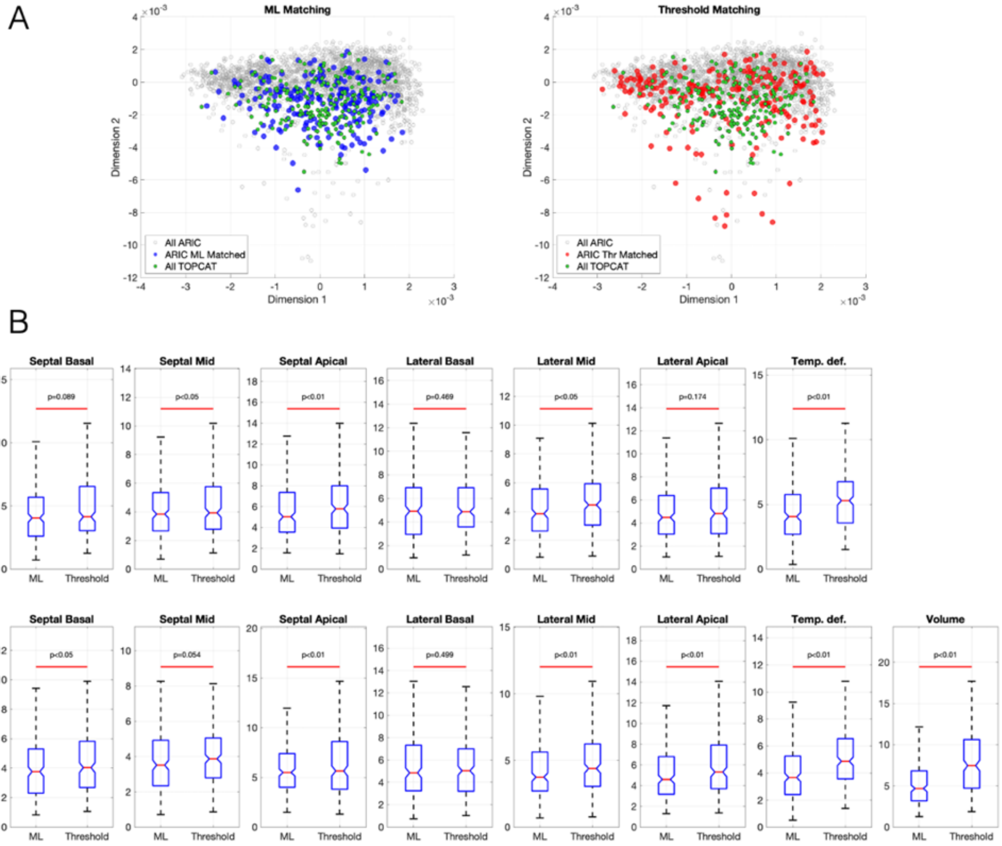


Figure 6 TOPCAT ML-based patient matching based on clinical characteristics and cardiac mechanics

A) 2-dimensional space from Figure 4. TOPCAT patients are shown in green, whereas 'ML-matched ARIC patients' are shown in blue and 'Threshold-matched ARIC patients' are shown in red. The comparison of clinical characteristics between these two patient groups is shown in Table S1. B) Distribution of distances for each feature for both matching approaches. Lower values indicate a larger similarity to the trial patients. A bar above each distribution displays the statistical significance of the difference between both distance distributions. As it can be observed, the volume/strain profiles show a much better overlap when using the ML based matching than the threshold based, signifying good cardiac mechanic-based matching using the ML

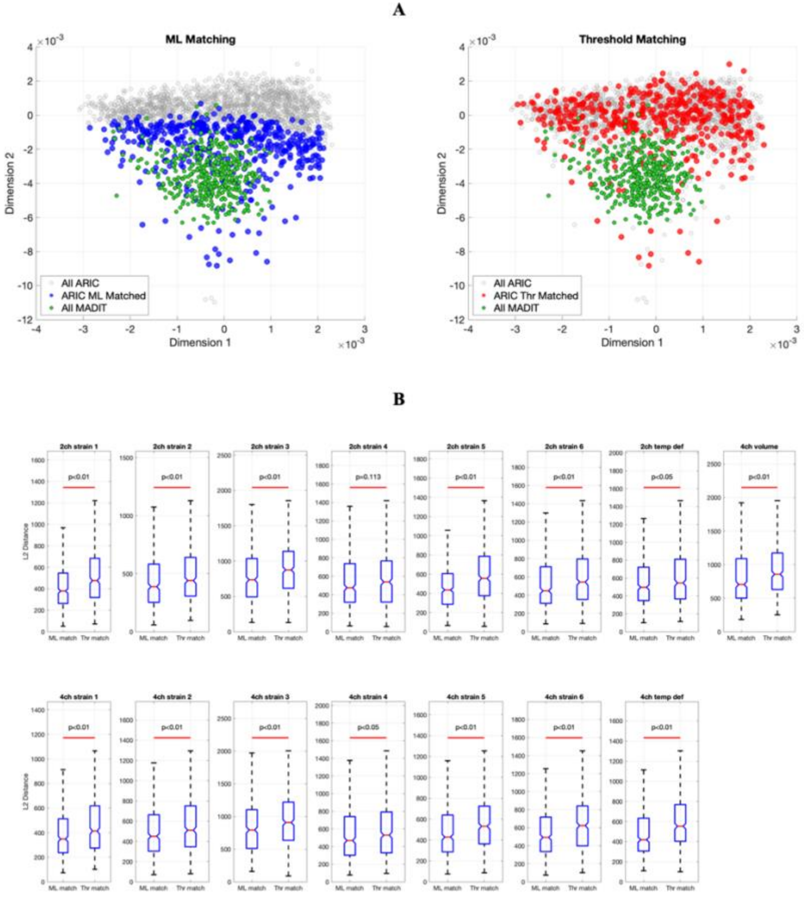


Figure 7 MADIT ML-based patient matching based on clinical characteristics and cardiac mechanics

A) 2-dimensional space from Figure 4. MADIT patients are shown in green, whereas 'ML-matched ARIC patients' are shown in blue and 'Threshold-matched ARIC patients' are shown in red. The comparison of clinical characteristics between these two patient groups is shown in Table S2. B) Distribution of distances for each feature for both matching approaches. Lower values indicate a larger similarity to the trial patients. A bar above each distribution displays the statistical significance of the difference between both distance distributions. As it can be observed, the volume/strain profiles show a much better overlap when using the ML based matching than the threshold based, signifying good cardiac- mechanic based matching using the ML.

1.4 Discussion

Our research demonstrates the capacity of unsupervised ML for analyzing large patient datasets, integrating diverse data types, and defining phenotypes associated with increased clinical risk (Figure 7). Besides providing insight in diverse phenotypes in the cohorts analyzed, this methodology can substantially improve the efficiency of clinical trials through one-to-one patient matching, reducing bias and expenses. Future work can enhance the ML-derived space resolution and explore real-world applications in clinical trials.

Unsupervised ML shows potential as a valuable data analysis tool to tackle the challenge of analyzing large patient datasets. It harnesses the capability to fuse data from heterogeneous sources (e.g., demographic characteristics, laboratory biomarkers, imaging) and reduce its dimensionality to create a low dimensional space that captures the salient characteristics of the cohort enabling direct comparisons between patients based on their integrated data conglomerate, exploration of physiologically sensible patterns (16), or evaluation of therapy response (17). The ability to integrate data types and capture data patterns can therefore enable the fusion of datasets with similar data profiles to directly compare different populations based on a desired input. Combining general population cohorts with clinical trials enables the projection of ‘high outcome’ patients into a ‘low outcome’ population to define phenotypes/regions related to increased clinical risk. In our example, the addition of HF patients to the ARIC population defined a region initially signaling intraventricular conduction abnormalities as the region grouping the majority of HFrEF patients, hence, the ARIC population in this ‘neighborhood’ was recognized as a specific phenotype with corroborated risks (Figure 5A). Furthermore, the fusion of trial datasets can help define a spectrum of disease. In our example, HFpEF, which is generally recognized as a heterogeneous clinical syndrome, can be seen distributed across the three phenoclusters, resonating the clinical challenge in predicting risk in these patients. Although participant selection is led by strict predefined inclusion and exclusion criteria, trials inevitably recruit patients that fall into different parts of the disease spectrum, potentially leading to challenges in the interpretation of results (29). Therefore, sub-phenotyping a diagnosis/cohort that is perceived homogeneous is of high clinical interest.

Finally, the proposed methodology bears the potential to increase the efficiency of running clinical trials. One-to-one matching of patients based on cardiac mechanics could be used to replace a percentage of the control arm with in-silico patients, resulting in a reduction of expenses and ‘recycling’ of general population clinical data.

Limitations of the materials used are adjacent to general limitations – data quality, cohort size, and the number of outcomes. As for methodological limitations, the most important relates to the fact that the currently used implementation of MKL has a high computational cost for big datasets (>1000 observations); thus, limiting the amount of patients used for training the representation. This could be addressed using online learning or Nyström approximation techniques (30,31). Lastly, when using this approach, one should bear in mind that the digital control database should be updated regularly to avoid confounders that affect large sections of the population i.e. the coronavirus epidemic, consequentially modifying the baseline event rate.

Future work will focus on how to enable integration of more complex datasets and data formats (e.g., unprocessed medical images, information from multiple cardiac beats, integration of exercise stress data) leading to increased 'resolution' of the ML-derived space, enabling more complex patient phenotyping through feature maps, regression analysis and clustering. ML matching could be tested to improve current propensity score matching techniques by integrating it into a real-world scenario to find controls for ongoing clinical trials. This would ultimately result in a high-quality win ratio (32) that allows for more accurate evaluation of the efficacy of a trial.

1.5 Conclusion

The presented analysis serves as a proof-of-concept for an unsupervised ML approach in analyzing registry and trial datasets. ML can enhance the integration of patient data and improve disease phenotyping, fuse different patient cohorts to build trust towards findings, and potentially increase the efficiency of running clinical trials through advanced matching of patients from the general population to those in trial control arms.

1.6 Supplementary Material

Table S 1 Clinical and echo characteristics of the TOPCAT matched general population controls, the matched ARICs, and the EF and HF history selected ARICs

	TOPCAT (n=218)	ML Matched ARICs (n=218)	Threshold Matched ARICs (n=218)	P-value$_1$ TOPCAT vs ML	P-value$_2$ TOPCAT vs Thr
Outcome, n (%)	52 (23.85%)	22 (10.09%)	20 (9.17%)	<0.001	<0.001
Age, years	71.19 (62.29 to 78.26)	75.12 (71.43 to 80.36)	75.78 (72.01 to 80.06)	<0.001	<0.001
Female sex, n (%)	115 (52.75%)	98 (44.95%)	106 (48.62%)	0.103	0.389
Caucasian, n (%)	169 (77.52%)	178 (81.65%)	175 (80.28%)	0.285	0.481
BMI, kg/m2	30.37 (26.57 to 34.93)	26.92 (24.45 to 29.97)	28.16 ± 5.16	<0.001	<0.001
Diabetes, n (%)	64 (29.36%)	66 (30.28%)	73 (33.49%)	0.834	0.353
Current smoker, n (%)	27 (12.39%)	16 (7.34%)	10 (4.59%)	0.077	0.003
Myocardial infarction, n (%)	56 (25.69%)	3 (1.38%)	4 (1.83%)	<0.001	<0.001
Prior heart failure, n (%)	132 (60.55%)	3 (1.38%)	16 (7.34%)	<0.001	<0.001
Blood pressure medication, n (%)	166 (76.15%)	166 (76.15%)	173 (79.36%)	1	0.420
Systolic blood pressure, mmHg	126.71 ± 15.23	131.4 ± 18.55	130.84 ± 19.02	0.004	0.0126
Heart rate, per minute	66 (60 to 74)	62 (57 to 68)	61 (55 to 66)	<0.001	<0.001
QRS width, ms	92.5 (84 to 108)	92 (84 to 102)	95.07 (86 to 108)	0.262	0.436
LVMi, g/m2	99.36 (82.59 to 119.36)	78.36 (67.13 to 89.24)	81.27 (69.25 to 97.73)	<0.001	<0.001
LVEDVi, ml/m2	47.19 (37.89 to 55.18)	43.63 (37.59 to 50.41)	45.52 (38.16 to 55)	0.007	0.651
LAVi, ml/m2 (IQR)	29.04 (21.7 to 35.09)	25.45 (20.83 to 31.36)	26.22 (20.95 to 31.92)	0.002	0.003
LVEF, %	60 (55 to 65)	64.45 (60.7 to 67.7)	57.45 (55.3 to 61.7)	<0.001	0.662
4CH GLS, %	-17.18 ± 3.11	-16.85 ± 2.2	-17.28 ± 2.44	0.204	0.692

IQR – interquartile range; BMI – body mass index; LVMi – LV mass indexed to body surface area; LVEDVi - LV end-diastolic volume indexed to body surface area; LAVi - left atrial volume indexed to body surface area, LVEF – LV ejection fraction; 4CH- 4-chamber; GLS – global longitudinal strain.

Table S 2 Clinical and echo characteristics of the MADIT matched general population controls, the matched ARICs, and the EF and HF history selected ARICs.

	MADIT (n=429)	**ML Matched ARICs (n=429)**	**Threshold Matched ARICs (n=429)**	**P-value$_1$ MADIT vs ML**	**P-value$_2$ MADIT vs Thr**
Outcome, n (%)	80 (18.65%)	47 (10.96%)	34 (7.93%)	0.0015	<0.001
Age, years	65 (57 to 71.25)	75.34 (71.61 to 80.43)	75.79 (71.69 to 80.37)	<0.001	<0.001
Female sex, n (%)	104 (24.24%)	181 (42.19%)	265 (61.77%)	<0.001	<0.001
Caucasian, n (%)	391 (91.14%)	350 (81.59%)	359 (83.68%)	<0.001	<0.001
BMI, kg/m2	28 (25.08 to 30.7)	27.78 (24.58 to 31.05)	27.43 (24.44 to 30.88)	0.595	0.272
Diabetes, n (%)	120 (27.97%)	136 (31.7%)	133 (31%)	0.232	0.330
Current smoker, n (%)	56 (13.05%)	33 (7.69%)	24 (5.59%)	0.010	<0.001
Myocardial infarction, n (%)	182 (42.42%)	7 (1.63%)	5 (1.17%)	<0.001	<0.001
Prior heart failure, n (%)	157 (36.6%)	10 (2.33%)	16 (3.73%)	<0.001	<0.001
Blood pressure medication, n (%)	411 (95.8%)	327 (76.22%)	325 (75.76%)	<0.001	<0.001
Systolic blood pressure, mmHg	120 (108 to 132)	131 (119 to 143)	130 (117 to 141)	<0.001	<0.001
Heart rate, per minute	63 (56 to 70)	63 (57 to 71)	61 (55 to 66.07)	0.261	0.012
QRS width, ms	158 (142 to 170)	92 (86 to 105)	90 (84 to 100)	<0.001	<0.001
LVMi, g/m2	106.82 ± 19.55	85.67 ± 23.68	80.2 ± 20.83	<0.001	<0.001
LVEDVi, ml/m2	120.01 (108.01 to 137.91)	44.65 (37.01 to 52.24)	41.91 (36.04 to 49.49)	<0.001	<0.001
LAVi, ml/m2 (IQR)	45.5 (39.76 to 52.62)	25.13 (20.12 to 32.47)	25.79 (20.89 to 31.11)	<0.001	<0.001

LVEF, %	25 (20 to 28)	62.6 (59.1 to 66.6)	68.8 (57.4 to 74.73)	<0.001	<0.001
4CH GLS, %	-9.45 (-11.08 to -7.78)	-16.33 (-17.43 to -14.99)	-18.76 (-20.55 to -17.09)	<0.001	<0.001

IQR – interquartile range; BMI – body mass index; LVMi – LV mass indexed to body surface area; LVEDVi - LV end-diastolic volume indexed to body surface area; LAVi - left atrial volume indexed to body surface area, LVEF – LV ejection fraction; 4CH- 4-chamber; GLS – global longitudinal strain.

2. CROSS-CULTURAL MACHINE LEARNING FROM ULTRASOUND AND CLINICAL CHARACTERISTICS FOR PREDICTING ADVERSE PERINATAL OUTCOMES

Abstract

a) Background

The occurrence of adverse perinatal outcomes remains a global critical challenge in both high- and low-income settings. Current practice provides no effective prevention or therapy, and predictive methods have shown limited performance. In this study we analyze two distinct cohorts of pregnant women at increased risk of adverse perinatal outcome and implement machine learning models to improve its prediction and discern their differential characteristics.

b) Methods

We considered two distinct patient cohorts to train and validate XGBoost classifier models to predict adverse perinatal outcomes from a comprehensive set of maternal and fetal characteristics including socio-demographic information, current pregnancy information, past pregnancy histories for the mothers, fetal biometry, and feto-placental Doppler measurements. The patient data were sourced from the IMPACT study in Barcelona, Spain, and the FeDoC study in a peri-urban settlement of Karachi, Pakistan. The data included pregnancies in the third trimester (>28 weeks). The models were trained on varying subsets of these features to evaluate the impact of different combinations on predictive ability. We also investigated the generalization of these models across the

two cohorts. Model performance metrics were evaluated, and interpretability was assessed using SHAP values, which quantify the impact of each feature on the model predictions.

c) Findings

Within the IMPACT dataset, biometrics stood out as a primary predictive variable set, with its combination with Doppler indices achieving the highest prediction accuracy.

Conversely, the FeDoC data revealed a higher information content in maternal clinical data over biometrics and Doppler, with the combination of clinical and Doppler data proving to be the best (with a marginal effect of biometrics).

Transfer learning scenarios indicated a consistent trend of biometrics as the most predictive, especially when paired with Doppler data. We present ROCs and SHAP values in Figure 1.

SHAP values identified the features exerting the most influence on model decisions, which exhibited distinct behaviors in within-dataset and transfer scenarios. The correlation between feature values and SGA decision remained consistent across scenarios.

relied heavily on biometrics and blood flow measurements along with the number of previous pregnancies.

Predictions of alternate outcomes had poor performance.

d) Interpretation

Our study obtained predictability results of around 80% AUC in high-income settings and around 70% AUC in low-income settings. The disparities between the cohorts highlight the role of demographic, socioeconomic, and healthcare factors in high-risk pregnancies. The differential predictability is attributed to different etiologies for SGA in the populations, with high-income cohorts leaning more on hemodynamic factors and low-income cohorts highlighting maternal undernutrition. The study showcases the potential of machine learning research in adverse perinatal event prediction, potentially enabling healthcare providers to identify and intervene in high-risk pregnancies, especially in regions with restricted healthcare resources.

2.1 Introduction

Despite a 53% global reduction in mortality among children under five, slow progress has been made in decreasing stillbirths and newborn deaths (33). These outcomes constitute a large proportion of deaths in children under five and include many infants who are classified as small for gestational age (SGA) - those born weighing below the 10th percentile (34). Apart from having a higher risk of perinatal mortality, SGA infants are at a higher risk of perinatal morbidity, such as poor neurodevelopment and increased cardiovascular problems as they grow (35,36). SGA fetuses may have varying associated risk factors, including placental abnormalities or insuficiency, maternal cardiovascular problems, malnutrition or protein deficits, iron deficiency and anemia (36).

Timely recognition with a comprehensive care pathway could prevent 75% of deaths among children under five (37). This care should span the prenatal, intrapartum, and postnatal period for both mother and child. This challenge is present worldwide but gets exacerbated in Low- or Middle-Income Countries (LMICs), like Pakistan, due to its limited healthcare resources. In consequence, Pakistan has been rated by UNICEF as the “riskiest place” for the birth of a child (8,38,39).

Several factors and biomarkers are known to have certain predictive value at the time of identifying at-risk pregnancies, such as socio-demographic information, current pregnancy health indicators, or past pregnancy histories (40). In addition, the fetal heart is known to adapt to adverse intrauterine environments, leading to compensatory mechanisms like cardiac remodeling and blood-flow redistribution across fetal vessels (41). Thus, besides its capacity to detect fetal anomalies (such as congenital heart disease), B-mode and Doppler echocardiography on the feto-placental anatomy and circulation has shown potential in predicting fetal growth restriction and neonatal mortality (42,43). Even with the known predictive power of these variables, our capacity to identify “at risk” fetuses exhibiting at-risk sub-clinical features is limited, emphasizing the need for technologically advanced solutions (44).

Emerging technologies such as Machine Learning (ML) hold the potential to revolutionize various fields. Its capacity to integrate and interpret large volumes of data, encompassing many samples and characteristics, is especially promising. This ability could be instrumental in identifying high-risk fetuses, thereby improving healthcare delivery and services, particularly in Low-and-Middle-Income Countries (LMICs). (17,45). Our primary objective was to develop machine learning models that can leverage comprehensive data collected to identify SGA fetuses from high-risk cohorts. Providing caregivers in resource-constrained settings with such advanced technologies would catalyze pregnancy risk stratification thus allowing timely interventions (46).

2.2 Methods

a) Study setting and participants

The first of the two studies included in this work, the IMPACT study **(**ClinicalTrials.gov Identifier: NCT03166332**)** is an randomized controlled clinical trial that took place at the Hospital Clinic of Barcelona (Spain) from 2017 to 2020. The 1221 participants were pregnant women, over 18 years old, who were at a high risk of having a growth-restricted fetus according to the Royal College of Obstetrics and Gynecology (RCOG) Guidelines (47). These women were randomly allocated to three arms of intervention: a Mediterranean diet, a mindfulness-based stress reduction program (MBSR), or no intervention. Details are described elsewhere (48).

The FeDoC (Fetal Doppler Consortium) study is a prospective observational cohort study that took place in Ibrahim Hyderi, Karachi, Pakistan in 2018 **(**ClinicalTrials.gov Identifier: NCT03398551)**.** The rationale and detailed study procedures have been described earlier (49). Participants included 694 pregnant women between 22-34 weeks of gestation. Both cohorts underwent standardized ultrasound studies during the study period.

b) Data selection and processing

Pregnancies with available data on feto-placental ultrasound at third trimester were included in this study. The data variables included in this study were those that were present in both cohorts and corresponding to pregnancies post 28 weeks. They comprised a total of five distinct categories: maternal socio-demographic information, past pregnancy history, current pregnancy health indicators, feto-placental Doppler and fetal biometry. Pulsatility indices for the umbilical artery and middle-cerebral artery were extracted from focused Doppler pulsed-wave acquisitions. Cerebro-placental ratio was calculated by dividing Middle Cerebral Artery PI by Umbilical Artery PI. Additionally, fetal biometry was obtained from routine ultrasound examination including biparietal diameter, head circumference, abdominal circumference, and femur length.

These measurements were acquired thrice and averaged. A comprehensive list is included in Table S3.

Observations missing continuous variables were removed from the analyses and those containing missing entries for binary variables were set to 0, as a conservative stance. Finally, to harmonize FeDoC and IMPACT datasets, high risk behaviors (smoking, sniffing/chewing tobacco, chewing betel nut, alcohol consumption, and drug use), education level and employment status variables were mapped. Details on the data harmonization are included in the supplementary material.

c) Outcomes and reference standards

The IMPACT and FeDoC studies primarily investigated SGA. FeDoC additionally assessed stillbirth or early neonatal mortality. These studies had secondary outcomes encompassing a wide array of adverse perinatal outcomes (APO), which included conditions such as preeclampsia, severe FGR, metabolic acidosis (in IMPACT), prematurity, birth asphyxia, neonatal sepsis, and low birth weight.

For this analysis, we focused our effort on the prediction of pregnancies labeled as SGA, defined as birthweight below the 10th centile according to the INTERGROWTH-21 standard (50). However, in Barcelona, we also identified SGA using a local Spanish standard developed in 2007 (51), which is tailored to regional characteristics. The differences when using differing methodologies are included in the supplement.

Preterm birth was defined as delivery prior to 37 weeks gestation (52).

d) Statistical analysis

Differences between the characteristics of the two studies were evaluated using a T-test or Mann-Whitney U test for continuous variables, depending on the normality of the data. The chi-squared test was used for binary variables.

e) Predictive models

To identify the most predictive factors for SGA, several supervised gradient-boosted ensemble models were developed using the XGBoost Python implementation (53). The models were initialized using a class-weighting strategy to account for class imbalance and used binary cross entropy as a measure of performance. They were trained using a version of repeated cross-validation. For this, in each of a total of 10 iterations, the two cohorts were partitioned as 70 train/30 test stratified by the interrogated outcome. Model training was done following Bayesian optimization for hyper-parameter tuning, coupled with 5-fold cross-validation. Area Under the Receiver Operating Characteristic Curve (AUROC) was used as performance metric of the cross-validation setting. It is worth mentioning that no feature selection strategy was implemented, but Lasso and Ridge regularization were implemented during model's training, which effectively penalize the model for using irrelevant or redundant features, thus reducing overfitting, and improving model's generalization.

Given the limited number of studies and the class-imbalance scenario, model's stability and performance was assessed through the average sensitivity, specificity, positive predictive value, negative predictive value (obtained at a 10% false-positive rate), and AUROC over the 10 iterations with their corresponding standard deviations. However, for further interpretation of results, we report the best model (out of 10 iterations) as we assume it to be the one that has best captured the underlying data patterns and relationships with the interrogated outcome. The best models were further interpreted using Shapley Additive Explanations (SHAP package for Python), to understand the contribution and influence of each variable on the model's predictions.

Given the different risk factors of SGA, we explored the training of the models with differing sets of relevant features, and their combinations, to analyze the differential power they brought to the prediction. Namely, the sets used were:

- Set 1: Clinical data (socio-demographics, ongoing pregnancy health indicators, and past pregnancy histories), to assess maternal characteristics.
- Set 2: Fetal biometry, to evaluate fetal size.
- Set 3: Feto-placental Doppler measurements, to assess (placenta-induced) fetal hemodynamic changes.
- Set 4: Clinical data + biometry.
- Set 5: Clinical data + Doppler.
- Set 6: Biometry + Doppler.
- Set 7: Clinical data + biometry + Doppler.

Finally, we investigated the generalization of these models across the two cohorts.

We also explored the predictive ability of these data for predicting low birthweight (defined as below 2.5kg (54), to assess potential problems with GA estimation), preterm births, and stillbirths or neonatal deaths (only present in FEDOC).

2.3 Results

A total of 746 women were included from the IMPACT trial and 520 from the FeDoC study. We show the characteristics of the included cohorts in Table 2.

Table 2 Maternal and perinatal characteristics of the study populations

Variables	IMPACT (n=746)	FeDoC (n=520)	p-value
Current pregnancy Health Indicators			
Maternal Age (years)	37.41 (34.47 to 40.50)	28.00 (23.00 to 30.00)	<0.001
Maternal Height (cm)	163.53 (±6.30)	154.83 (±5.67)	<0.001
Maternal Weight (kg)	72.80 (65.72 to 81.70)	57.50 (51.20 to 65.83)	<0.001
Maternal Body mass index (kg/m^2)	27.08 (24.80 to 30.32)	23.99 (21.82 to 27.34)	<0.001
Maternal Systolic Blood Pressure at the time of visit (mmHg)	108.00 (100.25 to 115.00)	107.00 (99.00 to 114.25)	0.0162
Maternal Diastolic Blood Pressure at the time of visit(mmHg)	71.00 (66.00 to 77.00)	70.00 (64.00 to 75.00)	0.0018
Maternal Hemoglobin Level (g/dL)	11.70 (11.10 to 12.30)	9.08 (8.07 to 10.05)	<0.001
Antenatal Care Access	746 (100.00%)	372 (71.54%)	<0.001
Morbidities			
Preeclampsia	83 (11.13%)	29 (5.58%)	0.0009
Eclampsia	1 (0.13%)	1 (0.19%)	1*
Gestational Diabetes Mellitus	67 (8.98%)	11 (2.12%)	<0.001
Anemia or Iron Deficiency	295 (39.54%)	234 (45.00%)	0.0603
Fever or Antibiotic Use	64 (8.58%)	167 (32.12%)	<0.001
Pregnancy-related Bleeding	34 (4.56%)	25 (4.81%)	0.9425
Maternal Socio-Demographics			
Work Status			
Unemployed	50 (6.70%)	507 (97.50%)	<0.001
Self-Employed	55 (7.37%)	7 (1.35%)	<0.001
Private/Student/Other Employment	641 (85.92%)	6 (1.15%)	<0.001

Education level			
No or Primary Education Level	27 (3.62%)	397 (76.35%)	<0.001
Secondary or Technology Education Level	219 (29.36%)	117 (22.50%)	0.0067
University Education Level	500 (67.02%)	6 (1.15%)	<0.001
Risk related habits			
No Health Risk Habits	437 (58.58%)	257 (49.42%)	0.0033
Stopped Risk Habits After Pregnancy Confirmation	256 (34.32%)	9 (1.73%)	<0.001
Continued Risk Habits During Pregnancy	53 (7.10%)	254 (48.85%)	<0.001
Past pregnancy histories			
History of Normal Previous Pregnancies	464 (62.20%)	424 (81.54%)	<0.001
History of Previous Preterm Births	36 (4.83%)	156 (30.00%)	<0.001
History of Previous Fetal Deaths	304 (40.75%)	108 (20.77%)	<0.001
Fetal growth			
Gestational Age at US (weeks)	33.30 (32.40 to 34.10)	31.21 (30.43 to 32.57)	<0.001
Gestational Age at clinical data collection (weeks)	35.00 (34.00 to 36.00)	31.14 (30.43 to 32.57)	<0.001
Estimated Fetal Weight percentile	46.28 (24.27 to 70.29)	34.64 (15.06 to 65.40)	<0.001
Head Circumference (cm)	30.30 (29.30 to 31.20)	28.65 (27.95 to 29.76)	<0.001
Abdominal Circumference (cm)	29.20 (28.00 to 30.50)	27.52 (26.54 to 28.96)	<0.001
Biparietal Diameter (cm)	8.30 (8.00 to 8.60)	8.01 (7.77 to 8.30)	<0.001
Femur Length (cm)	6.30 (6.10 to 6.50)	6.13 (5.88 to 6.40)	<0.001
Estimated Fetal Weight (g)	2109.56 (1886.45 to 2340.53)	1786.91 (1630.30 to 2078.55)	<0.001

Feto-placental Blood Flow Measurements			
Middle Cerebral Artery Pulsatility Index	1.89 (1.68 to 2.13)	1.85 (1.58 to 2.15)	0.0911
Umbilical Artery Pulsatility Index	0.94 (0.84 to 1.06)	1.11 (0.98 to 1.25)	<0.001
Cerebro-Placental Ratio	2.00 (1.71 to 2.35)	1.67 (1.38 to 2.01)	<0.001
Outcomes or Interventions			
Small for Gestational Age (IG21)	78 (10.46%)	87 (16.73%)	0.0011
Neonatal Death	0 (0.00%)	33 (6.35%)	<0.001
Preterm	24 (3.22%)	141 (27.12%)	<0.001
C-section	242 (32.44%)	93 (17.88%)	<0.001

Values are reported as mean and standard deviation for normally distributed continuous variables and as median and IQR for non-normally distributed continuous variables. Binary variables are presented as counts and percentages. Risk related habits encompass alcohol consumption and drug use. *Yates correction produced a p-value of 1.

In Table 3 we show the mean and standard deviation, resulting from 10 launches, when training with each feature set and with each combination of within-dataset and transfer scenarios. Within the IMPACT dataset, biometrics stood out as a primary predictive variable set, with its combination with Doppler indices achieving the highest prediction accuracy. Conversely, the FeDoC data revealed a higher information content in maternal clinical data over biometrics and Doppler, with the combination of clinical and Doppler data proving to be the best (with a marginal effect of biometrics). Transfer learning scenarios indicated a consistent trend of biometrics as the most predictive, especially when paired with Doppler data. We present ROCs and SHAP values in Figure 8.

Table 3 SGA prediction AUCs for the different sets of features

	Set 1	**Set 2**	**Set 3**	**Set 4**	**Set 5**	**Set 6**	**Set 7**
	Clinical	Biometrics	Doppler	Clinical Biometrics	Clinical Doppler	Biometrics Doppler	Clinical Biometrics Doppler
IMPACT – IMPACT	59.4 ± 5.3	78.7 ± 2.0	65.8 ± 4.7	75.0 ± 3.4	67.4 ± 5.8	80.3 ± 3.0	80.3 ± 2.4
FeDoC – FEDOC	69.9 ± 3.9	68.5 ± 2.8	68.2 ± 5.0	70.6 ± 4.9	73.0 ± 4.0	69.4 ± 3.4	73.1 ± 3.6
IMPACT – FEDOC	56.3 ± 5.9	67.3 ± 5.5	61.3 ± 3.8	64.3 ± 6.1	63.6 ± 4.6	70.3 ± 4.5	71.0 ± 6.5
FeDoC – IMPACT	58.0 ± 6.2	70.1 ± 4.4	64.0 ± 4.7	68.8 ± 4.9	64.4 ± 3.3	76.8 ± 3.5	68.4 ± 6.7

SHAP values identified the features exerting the most influence on model decisions, which exhibited distinct behaviors in within-dataset and transfer scenarios. The correlation between feature values and SGA decision remained consistent across scenarios. Specifically, fetal biometry, MCA PI, CPR, maternal height and weight, previous pregnancies, and previous preterm pregnancies were negatively correlated, whereas gestational age at ultrasound visit, UA PI, and maternal hemoglobin levels were positively correlated with SGA classification.

When trained and tested within the IMPACT dataset, the model deemed the gestational age at ultrasound visit and biometric measurements of the fetus as the most salient features, followed by Doppler measurements (CPR and MCA), maternal weight and hemoglobin levels. However, when the IMPACT-trained model was tested on the FeDoC dataset, the gestational age at ultrasound emerged as the most significant feature, alongside biometric measurements of abdominal circumference, biparietal diameter, and femur length. The model also attributed importance to Doppler measures and maternal weight in its predictions.

In contrast, when trained and tested within the FeDoC dataset, the model prioritized CPR and UA PI as the most significant features, additionally considering maternal height, fetal head circumference, history of previous normal pregnancies and preterm births. Yet, when this FEDOC-trained model was tested within the IMPACT dataset, it relied heavily on biometrics and blood flow measurements along with the number of previous pregnancies.

We add a detailed report of the results interrogating alternate outcomes in the supplement. In a nutshell, low birthweight had moderate to good predictions both within-dataset and in transfer learning cases. Prediction of preterm deliveries had a moderate performance within IMPACT but fell to random chance if we excluded those due to C-section. The rest of outcome predictions had poor performance.

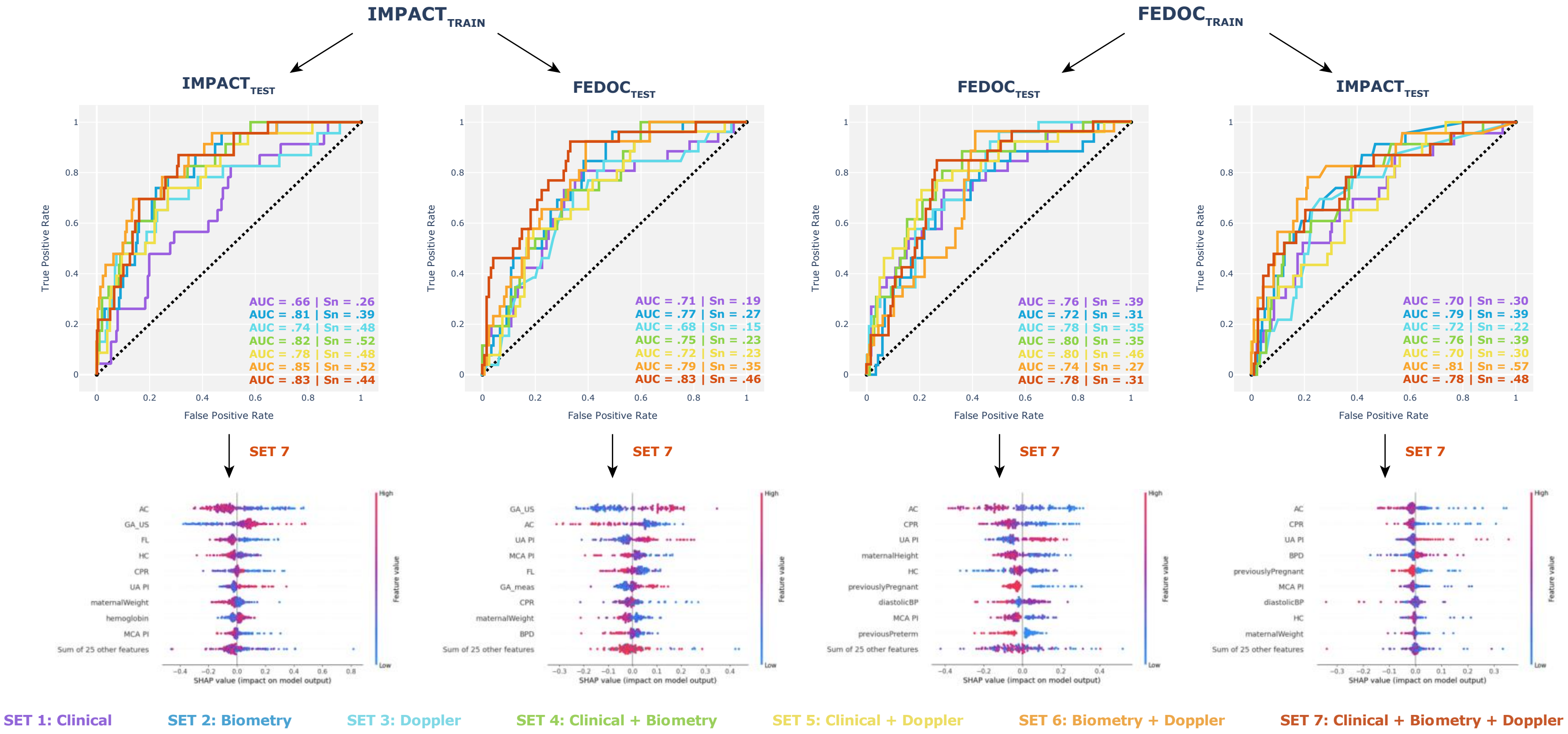


Figure 8 Performance metrics and feature interpretation of the models trained.

In the first row, we show ROCs trained on every feature set and tested within and across datasets. Below, we reported the SHAP values corresponding to the model with best performance using the full feature set. Sn, Sensitivity; HC, Head Circumference; AC, Abdominal Circumference; FL, Femur Length; BPD, BiParietal Diameter; GA_US, Gestational Age at Ultrasound scan; UA PI, Umbilical Artery Pulsatility Index; MCA PI, Middle Cerebral Artery Pulsatility Index; CPR, Cerebro-Placental Ratio.

2.4 Discussion

This study, to our knowledge, is pioneering in the use and comparison of ML models trained on an array of materno-fetal features to predict SGA fetuses. Moreover, we investigate their predictive performance in both a high-risk high-income population and a LMIC setting, investigating potential different causes for increased risk as well as evaluating transferability of ML-models across these two settings.

The disparities observed across the cohorts underline the influence of demographic, socioeconomic, and healthcare factors on high-risk pregnancies and maternal health outcomes. Understanding these differential baseline characteristics is crucial when interpreting and applying research findings across diverse settings and populations.

When training/testing in a similar setting, we obtained predictability results for SGA in the order of 80% AUC in IMPACT, which is comparable, or even slightly superior to current clinical practice while in FEDOC, we obtain values in the order of 70% AUC.

Two factors might contribute to the different performance, first, the etiologies for SGA and pregnancy factors that lead to it, can be different in the populations. For example, based on the results obtained, SGA in both cohorts could be attributed to different disease processes. Namely, in Barcelona, most cases appear to be associated to hemodynamic factors (likely due to placental insufficiency), given that when models trained on IMPACT, they mostly rely on fetal Doppler to forecast SGA. However, in FEDOC, the information gain from clinical (maternal) data is more relevant, pointing to risk factors linked to maternal undernutrition, as models trained on FeDoC weighted maternal characteristics more heavily. Secondly, a factor that might contribute to the different performance in both setting is related to the difference in GA at scan. By design, the GA at scan for the IMPACT study was 2 weeks later as compared to FeDoC (resulting in a mean GA of 33.3 vs 31.21 weeks). Especially when late SGA is involved, its prediction is known to improve with later GAs. From our analysis of the subgroup of GA 31-33 weeks (see S8), where there is the most overlap between the two cohorts, we observe a very similar AUC for IMPACT with a slight reduction in FeDoC (potentially attributed to the reduction in sample size), supporting the hypothesis of differences in dominant etiologies. For the transfer use of the models, from IMPACT to FeDoC there is little change in predictive power, while there seems to be a slight improvement in prediction in the transfer from FeDoC to IMPACT likely because there is a better GA match between the two, making the models more generalizable amongst the etiologies.

Another interesting aspect is the value of fetal biometry for the prediction of SGA. While it would seem straightforward that fetal size at scan should contribute to predicting fetal weight at birth, as is confirmed in IMPACT, in FeDoC it seems to contribute little on top of clinical information. This again might be related to two factors. First that SGA is more related to high (maternal) clinical risk rather than true placental disease. Here it might be that the traditional biometric markers, combining bone sizes and abdominal circumference, might not fully capture the effect of for example maternal malnutrition or other pregnancy problems. Secondly, the estimation of the GA at scan might be too inaccurate for biometry to truly reflect fetal growth. In FEDOC, as compared to IMPACT, the estimation of GA is much more challenging given lack of accurate information of the last menstrual cycle nor first trimester ultrasound.

When comparing our work to existing bibliography, results can vary greatly depending on the characteristics of the study populations, parameters included, or standards used to define SGA. For instance, the GA at data collection, origin, and size of the cohort as well as the complexity of ML models used to make the predictions and the variables collected to train them. The predictive ability of several features in our work were also of importance in previous studies (55–57). However, none of them focused on comparing high income and LMIC settings.

Among the limitations of our study, the comparison of two distinct cohorts to evaluate feature importance for SGA prediction may compromise accuracy due to intrinsically different decision boundaries. Moreover, the prediction of outcomes such as preterm pregnancies and perinatal death was constrained by class imbalance across both cohorts and varied rates of occurrence, indicating our cohorts might lack the necessary power to substantiate the predictive potential of these data.

In future research, larger datasets would help capture more preterm and perinatal death events, potentially providing insights into their predictability, which we were not able to assess. Also, longitudinal models that follow the growth of the fetus over time could be key to understand the gestation process and improve the prediction of SGA (58).

In summary, this paper assesses the predictability of SGA in developed and low-resource scenarios, analyzes the transferability of the models across them, and interprets model decisions considering specific input variables. Our findings allow for new insights into the pathophysiological role of the different descriptors from their use in the ML model and provides models for SGA prediction, to be validated in larger independent cohorts, thus, showcasing the use of ML to tackle heterogeneous information and cohorts, ultimately showing the potential of ML research in adverse perinatal event prediction. We recognize the complexity of perinatal mortality as a systemic issue compounded by multiple factors that resists straightforward solutions. However, given the pressing need to reduce disparities to improve global maternal and child health—especially in regions with restricted healthcare resources such as Pakistan—our findings can be seen as progress. They could equip caregivers with decision-support tools to identify and intervene in high-risk pregnancies, potentially restoring conditions to normalcy on time.

2.5 Supplementary Material

Table S 3 Features used for model training with their corresponding data type and imputation method.

Feature category	Feature name	Data type	Number missing
Ongoing pregnancy health indicators	Antenatal care	Binary	-
	High blood pressure	Binary	13
	Convulsions	Binary	4
	Bleeding	Binary	-
	Gestational Diabetes Mellitus	Binary	4
	Anemia or Iron Deficiency	Binary	2
	Maternal Hemoglobin	Continuous	44
	Fever or antibiotics	Binary	-
	Maternal Weight	Continuous	5
	Maternal Height	Continuous	-
	Systolic BP	Continuous	33
	Diastolic BP	Continuous	33

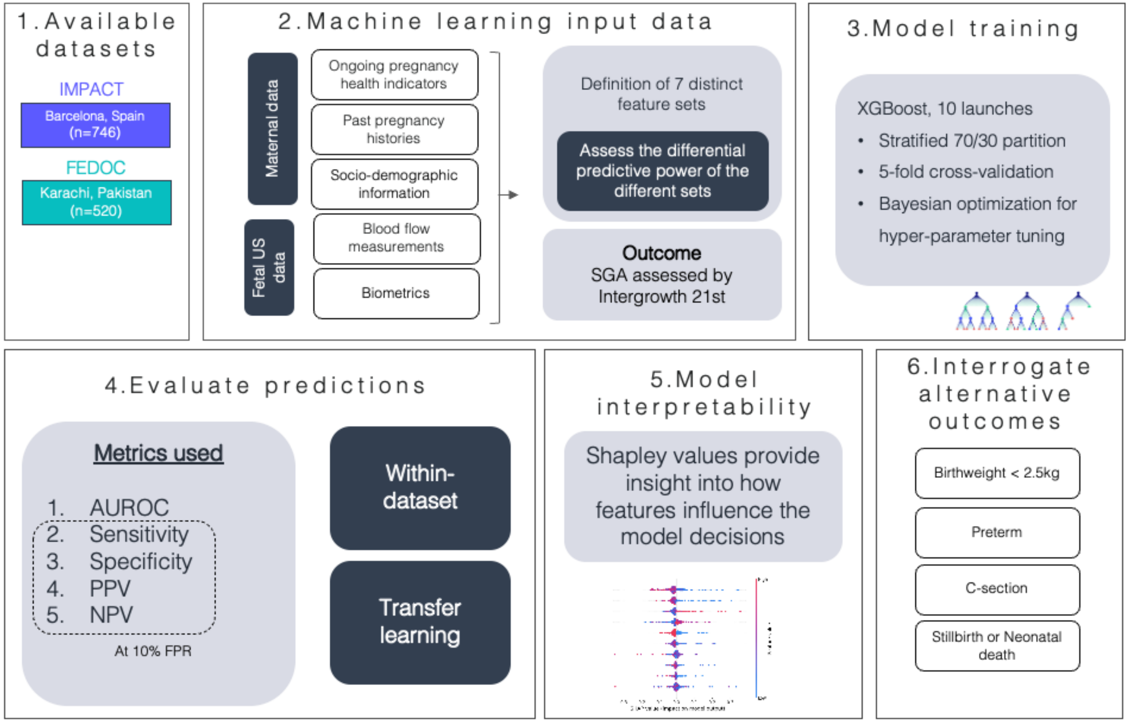
	Gestational age at US	Continuous	-
	Gestational age (measures)	Continuous	-
Fetal blood flow measurements	Umbilical artery PI	Continuous	-
	Mid cerebral artery PI	Continuous	-
	Cerebroplacental ratio	Continuous	-
Fetal size and development metrics	Head circumference	Continuous	-
	Abdominal circumference	Continuous	-
	Biparietal diameter	Continuous	-
	Femur Length	Continuous	-
	Gestational age	Continuous	-
Past pregnancy histories	Previously pregnant	Binary	-
	Previous preterm	Binary	-
	Previous fetal death	Binary	-
	Age	Continuous	-

Maternal socio-demographic information	Education level	Binary (one hot encoded into categories)	-
	Work status	Binary	-
	Risk habits	Binary	-
Outcomes	SGA	Binary	
	Preterm	Binary	2
	Neonatal death	Binary	2
	C-section	Binary	2
	Birth Weight < 2.5 kg	Binary	38

Data processing

In the FeDoC dataset, smoking, sniffing/chewing tobacco, and chewing betel nut were recoded into values of 0, 1, and 2, representing non-usage, cessation, and ongoing usage respectively. Likewise, in the IMPACT dataset, habits like alcohol consumption, drug use, and smoking were recoded into the same values. Each patient's risk score was then calculated as the maximum value among these risk scores. The education data from FeDoC was restructured into categories: 'no/primary' for 0-6 years of education, 'secondary/technology' for 7-12 years, and 'university' for more than 12 years, to correspond with the categories of the IMPACT survey. Similarly, FEDOC's employment categories, which originally included 'Does not work', 'Private Job', 'Other work', 'Self-employed', 'Employed', 'Midwife', and 'Student', were simplified to align with those of IMPACT, namely 'Unemployed', 'Private/Other/Student', and ‘Self-employed’.

Methodology diagram



1. Available datasets

IMPACT

Barcelona, Spain (n=746)

FEDOC

Karachi, Pakistan (n=520)

2. Machine learning input data

Maternal data

Ongoing pregnancy health indicators

Past pregnancy histories

Socio-demographic information

Fetal US data

Blood flow measurements

Biometrics

Definition of 7 distinct feature sets

Assess the differential predictive power of the different sets

Outcome
SGA assessed by Intergrowth 21st

3. Model training

XGBoost, 10 launches

- Stratified 70/30 partition
- 5-fold cross-validation
- Bayesian optimization for hyper-parameter tuning

4. Evaluate predictions

Metrics used

1. AUROC
2. Sensitivity
3. Specificity
4. PPV
5. NPV

At 10% FPR

Within-dataset

Transfer learning

5. Model interpretability

Shapley values provide insight into how features influence the model decisions

6. Interrogate alternative outcomes

Birthweight < 2.5kg

Preterm

C-section

Stillbirth or Neonatal death

Figure S 1 Schematic of the methodology followed in this work.

Results for other Outcomes (C-section, preterm, lbw, etc)

Table S 4 AUCs for prediction of alternative outcomes and all feature sets

		SGA (Intergrowth)	Birthweight < 2.5 kg	Preterm	C-section	Preterm (excluding C-section)	Stillbirth & neonatal death
IMPACT **vs** **IMPACT**	Set1	59.4 ± 5.3	69.2 ± 8.8	75.5 ± 6.3	65.2 ± 2.9	57.5 ± 7.7	–
	Set2	78.7 ± 2.0	74.5 ± 3.8	56.5 ± 4.5	55.2 ± 3.7	45.4 ± 17.1	–
	Set3	65.8 ± 4.7	70.1 ± 6.6	63.5 ± 10.2	48.7 ± 2.6	55.6 ± 7.0	–
	Set4	75.0 ± 3.4	76.7 ± 4.7	73.7 ± 8.0	65.2 ± 3.0	52.7 ± 11.5	–
	Set5	67.4 ± 5.8	70.8 ± 5.7	79.5 ± 7.4	63.4 ± 2.4	48.9 ± 10.6	–
	Set6	80.3 ± 3.0	78.1 ± 3.3	59.8 ± 9.6	52.1 ± 3.3	60.1 ± 10.5	–
	Set7	80.3 ± 2.4	80.6 ± 4.1	75.1 ± 7.5	64.8 ± 2.2	50.6 ± 14.4	–
FEDOC **vs** **FEDOC**	Set1	69.9 ± 3.9	65.0 ± 5.1	57.4 ± 4.7	63.0 ± 5.2	59.5 ± 3.5	46.6 ± 9.8
	Set2	68.5 ± 2.8	62.5 ± 3.9	55.2 ± 4.8	49.1 ± 2.7	50.5 ± 3.6	49.7 ± 8.3
	Set3	68.2 ± 5.0	66.8 ± 4.3	58.1 ± 3.6	48.6 ± 4.6	55.0 ± 4.7	46.8 ± 6.8
	Set4	70.6 ± 4.9	66.8 ± 4.2	57.7 ± 3.4	62.5 ± 5.3	59.5 ± 3.8	50.6 ± 9.9
	Set5	73.0 ± 4.0	70.8 ± 3.9	58.2 ± 4.2	63.0 ± 4.8	63.6 ± 2.6	45.8 ± 7.0
	Set6	69.4 ± 3.4	67.2 ± 5.8	59.0 ± 3.9	46.4 ± 4.8	55.5 ± 3.4	46.5 ± 6.9
	Set7	73.1 ± 3.6	70.2 ± 5.3	59.8 ± 4.1	61.4 ± 3.7	61.0 ± 3.9	46.0 ± 9.6
IMPACT **vs** **FEDOC**	Set1	56.3 ± 5.9	56.8 ± 6.7	50.4 ± 5.8	52.3 ± 5.0	48.3 ± 5.4	–
	Set2	67.3 ± 5.5	62.0 ± 5.3	47.0 ± 3.5	50.3 ± 5.0	50.4 ± 4.1	–
	Set3	61.3 ± 3.8	69.5 ± 5.2	54.8 ± 3.6	52.1 ± 4.3	57.2 ± 4.5	–
	Set4	64.3 ± 6.1	61.2 ± 5.0	50.2 ± 3.9	50.1 ± 8.1	48.0 ± 5.6	–
	Set5	63.6 ± 4.6	67.7 ± 7.5	53.2 ± 4.3	57.3 ± 5.2	55.4 ± 7.2	–
	Set6	70.3 ± 4.5	71.2 ± 5.1	53.6 ± 4.5	54.5 ± 2.6	56.7 ± 4.5	–
	Set7	71.0 ± 6.5	69.3 ± 5.5	53.2 ± 5.4	54.1 ± 6.0	55.4 ± 5.1	–

FEDOC vs IMPACT	Set1	58.0 ± 6.2	59.2 ± 5.6	49.6 ± 6.9	56.8 ± 6.5	41.0 ± 16.4	–
	Set2	70.1 ± 4.4	63.4 ± 10.9	51.1 ± 8.3	53.3 ± 2.9	44.6 ± 18.0	–
	Set3	64.0 ± 4.7	73.2 ± 6.5	58.5 ± 13.3	50.2 ± 1.9	66.1 ± 19.9	–
	Set4	68.8 ± 4.9	58.2 ± 5.0	53.4 ± 5.7	56.5 ± 3.9	41.7 ± 18.0	–
	Set5	64.4 ± 3.3	66.8 ± 7.1	52.3 ± 10.1	57.4 ± 4.4	51.3 ± 24.4	–
	Set6	76.8 ± 3.5	74.9 ± 5.8	56.7 ± 12.3	52.6 ± 4.5	67.5 ± 16.3	–
	Set7	68.4 ± 6.7	71.9 ± 8.0	61.4 ± 11.6	56.6 ± 3.4	46.6 ± 20.6	–

Table S 5 Comprehensive performance metrics for low birthweight (<2.5kg) prediction across all feature sets

		AUC Train	AUC Test	Sn	Sp	PPV	NPV
Impact vs Impact	Set1	74.6 ± 3.0	69.2 ± 8.8	30.7 ± 12.4	91.0 ± 2.0	17.8 ± 4.7	95.2 ± 0.7
	Set2	78.5 ± 1.2	74.5 ± 3.8	36.4 ± 9.8	90.4 ± 0.7	20.0 ± 4.3	95.5 ± 0.7
	Set3	70.9 ± 2.6	70.1 ± 6.6	28.6 ± 12.4	90.8 ± 1.1	16.5 ± 5.5	95.0 ± 0.8
	Set4	81.4 ± 2.5	76.7 ± 4.7	37.9 ± 10.1	90.3 ± 0.7	20.3 ± 4.3	95.6 ± 0.6
	Set5	75.5 ± 3.1	70.8 ± 5.7	43.6 ± 8.7	90.6 ± 0.8	23.5 ± 3.9	96.0 ± 0.6
	Set6	81.4 ± 1.6	78.1 ± 3.3	39.3 ± 13.3	90.5 ± 0.8	21.4 ± 6.4	95.7 ± 0.9
	Set7	83.0 ± 2.0	80.6 ± 4.1	47.9 ± 9.0	90.2 ± 0.8	24.5 ± 4.2	96.3 ± 0.6
FEDOC vs FEDOC	Set1	69.3 ± 2.3	65.0 ± 5.1	21.4 ± 6.6	90.4 ± 0.6	33.8 ± 6.5	82.9 ± 1.2
	Set2	65.4 ± 2.5	62.5 ± 3.9	10.4 ± 7.4	93.6 ± 4.2	nan ± nan	81.5 ± 1.2
	Set3	70.5 ± 2.0	66.8 ± 4.3	26.8 ± 7.0	90.8 ± 1.0	40.5 ± 7.6	84.0 ± 1.3
	Set4	70.8 ± 2.4	66.8 ± 4.2	20.7 ± 4.5	90.5 ± 0.6	33.8 ± 5.3	82.8 ± 0.8
	Set5	74.1 ± 2.1	70.8 ± 3.9	27.9 ± 5.5	90.5 ± 0.5	40.7 ± 3.7	84.1 ± 1.0
	Set6	70.8 ± 1.8	67.2 ± 5.8	28.2 ± 7.9	91.2 ± 1.2	42.3 ± 6.4	84.3 ± 1.4
	Set7	75.1 ± 2.2	70.2 ± 5.3	26.8 ± 8.6	90.3 ± 1.5	38.7 ± 6.9	83.9 ± 1.5
Impact vs FEDOC	Set1	74.6 ± 3.0	56.8 ± 6.7	13.9 ± 10.4	91.1 ± 1.4	23.4 ± 13.9	81.7 ± 1.7
	Set2	78.5 ± 1.2	62.0 ± 5.3	16.1 ± 10.3	91.3 ± 2.7	26.1 ± 12.8	82.1 ± 1.6
	Set3	70.9 ± 2.6	69.5 ± 5.2	25.7 ± 12.5	91.5 ± 2.9	nan ± nan	83.9 ± 2.0
	Set4	81.4 ± 2.5	61.2 ± 5.0	16.1 ± 9.5	91.2 ± 0.7	28.2 ± 13.5	82.1 ± 1.7
	Set5	75.5 ± 3.1	67.7 ± 7.5	17.1 ± 9.2	90.6 ± 0.8	28.4 ± 8.8	82.2 ± 1.6

	Set6	81.4 ± 1.6	71.2 ± 5.1	30.4 ± 8.2	90.3 ± 0.9	42.0 ± 7.4	84.6 ± 1.5
	Set7	83.0 ± 2.0	69.3 ± 5.5	26.1 ± 11.1	90.6 ± 0.7	37.8 ± 11.1	83.8 ± 2.0
FEDOC vs Impact	Set1	69.3 ± 2.3	59.2 ± 5.6	20.7 ± 9.3	90.9 ± 1.1	13.0 ± 5.5	94.5 ± 0.6
	Set2	65.4 ± 2.5	63.4 ± 10.9	22.1 ± 10.4	91.2 ± 1.2	14.3 ± 6.1	94.6 ± 0.7
	Set3	70.5 ± 2.0	73.2 ± 6.5	38.6 ± 9.1	90.4 ± 0.6	20.8 ± 3.9	95.7 ± 0.6
	Set4	70.8 ± 2.4	58.2 ± 5.0	15.0 ± 8.1	91.3 ± 1.4	10.4 ± 6.2	94.2 ± 0.5
	Set5	74.1 ± 2.1	66.8 ± 7.1	35.0 ± 13.0	91.3 ± 1.2	21.1 ± 7.3	95.5 ± 0.9
	Set6	70.8 ± 1.8	74.9 ± 5.8	41.4 ± 11.4	90.7 ± 0.9	22.6 ± 4.6	95.9 ± 0.7
	Set7	75.1 ± 2.2	71.9 ± 8.0	43.6 ± 11.3	90.4 ± 0.4	23.0 ± 4.9	96.0 ± 0.8

Table S 6 Comprehensive performance metrics for preterm prediction across all feature sets

		AUC Train	AUC Test	Sn	Sp	PPV	NPV
Impact vs Impact	Set1	77.9 ± 4.6	75.5 ± 6.3	37.2 ± 15.9	91.0 ± 1.3	11.6 ± 4.6	97.8 ± 0.5
	Set2	63.8 ± 5.6	56.5 ± 4.5	4.3 ± 6.6	92.0 ± 2.0	1.5 ± 2.3	96.8 ± 0.2
	Set3	65.6 ± 4.8	63.5 ± 10.2	18.6 ± 15.7	91.2 ± 2.4	5.6 ± 4.7	97.2 ± 0.5
	Set4	75.3 ± 4.6	73.7 ± 8.0	30.0 ± 18.6	91.5 ± 2.9	nan ± nan	97.6 ± 0.6
	Set5	76.4 ± 3.7	79.5 ± 7.4	45.7 ± 21.0	90.8 ± 0.7	13.4 ± 5.2	98.1 ± 0.7
	Set6	66.4 ± 5.7	59.8 ± 9.6	14.3 ± 11.1	92.3 ± 1.7	5.3 ± 3.6	97.1 ± 0.3
	Set7	75.3 ± 3.2	75.1 ± 7.5	30.0 ± 19.6	91.0 ± 1.0	9.2 ± 5.2	97.6 ± 0.7
FEDOC vs FEDOC	Set1	60.1 ± 2.9	57.4 ± 4.7	14.0 ± 5.7	90.9 ± 1.9	35.7 ± 8.0	74.2 ± 1.2
	Set2	59.6 ± 2.5	55.2 ± 4.8	11.4 ± 6.8	92.7 ± 3.7	nan ± nan	74.0 ± 0.9
	Set3	61.8 ± 1.1	58.1 ± 3.6	19.5 ± 7.5	91.0 ± 2.9	45.9 ± 12.1	75.5 ± 1.5
	Set4	61.7 ± 3.2	57.7 ± 3.4	15.0 ± 4.1	90.5 ± 0.8	36.3 ± 6.3	74.3 ± 0.9
	Set5	63.2 ± 2.3	58.2 ± 4.2	16.9 ± 5.3	90.8 ± 0.7	39.5 ± 8.5	74.8 ± 1.2
	Set6	62.6 ± 1.9	59.0 ± 3.9	15.0 ± 6.2	91.4 ± 3.0	nan ± nan	74.5 ± 1.1
	Set7	63.6 ± 2.1	59.8 ± 4.1	14.5 ± 5.6	90.5 ± 0.7	34.9 ± 8.8	74.2 ± 1.3
Impact vs FEDOC	Set1	77.9 ± 4.6	50.4 ± 5.8	12.1 ± 4.8	90.6 ± 0.5	31.1 ± 8.7	73.7 ± 1.1
	Set2	63.8 ± 5.6	47.0 ± 3.5	7.4 ± 3.8	91.0 ± 1.1	22.4 ± 8.4	72.7 ± 0.8
	Set3	65.6 ± 4.8	54.8 ± 3.6	12.9 ± 4.8	91.1 ± 0.9	33.8 ± 8.4	74.0 ± 1.1
	Set4	75.3 ± 4.6	50.2 ± 3.9	9.8 ± 5.1	92.0 ± 2.9	27.8 ± 13.4	73.4 ± 0.9
	Set5	76.4 ± 3.7	53.2 ± 4.3	14.3 ± 6.5	90.7 ± 1.0	34.1 ± 9.9	74.2 ± 1.4

	Set6	66.4 ± 5.7	53.6 ± 4.5	11.9 ± 6.8	91.1 ± 2.8	28.7 ± 13.5	73.7 ± 1.3
	Set7	75.3 ± 3.2	53.2 ± 5.4	13.8 ± 4.9	90.6 ± 0.5	34.1 ± 7.9	74.1 ± 1.1
FEDOC vs Impact	Set1	60.1 ± 2.9	49.6 ± 6.9	4.3 ± 6.6	93.2 ± 3.1	1.6 ± 2.5	96.8 ± 0.2
	Set2	59.6 ± 2.5	51.1 ± 8.3	7.2 ± 7.2	92.7 ± 3.0	nan ± nan	96.9 ± 0.3
	Set3	61.8 ± 1.1	58.5 ± 13.3	27.2 ± 13.5	91.4 ± 1.3	9.4 ± 5.0	97.5 ± 0.4
	Set4	61.7 ± 3.2	53.4 ± 5.7	11.4 ± 12.5	92.4 ± 2.8	3.6 ± 3.8	97.0 ± 0.3
	Set5	63.2 ± 2.3	52.3 ± 10.1	20.0 ± 13.1	91.8 ± 2.3	6.7 ± 3.7	97.3 ± 0.4
	Set6	62.6 ± 1.9	56.7 ± 12.3	18.6 ± 15.7	93.4 ± 3.1	nan ± nan	97.3 ± 0.5
	Set7	63.6 ± 2.1	61.4 ± 11.6	24.3 ± 21.2	92.0 ± 2.2	7.2 ± 6.1	97.4 ± 0.7

Table S 7 Comprehensive performance metrics for preterm excluding C-sections prediction across all feature sets

		AUC Train	AUC Test	Sn	Sp	PPV	NPV
Impact vs Impact	Set1	74.1 ± 6.7	57.5 ± 7.7	6.7 ± 13.3	94.5 ± 3.0	0.9 ± 1.7	98.7 ± 0.2
	Set2	65.5 ± 5.3	45.4 ± 17.1	10.0 ± 15.3	91.1 ± 1.7	1.3 ± 2.1	98.7 ± 0.2
	Set3	79.0 ± 6.1	55.6 ± 7.0	10.0 ± 15.3	91.6 ± 2.3	1.4 ± 2.1	98.7 ± 0.2
	Set4	69.9 ± 9.1	52.7 ± 11.5	20.0 ± 26.7	92.8 ± 2.9	2.7 ± 3.6	98.8 ± 0.4
	Set5	78.7 ± 7.9	48.9 ± 10.6	0.0 ± 0.0	95.0 ± 3.0	nan ± nan	98.6 ± 0.1
	Set6	76.0 ± 5.8	60.1 ± 10.5	10.0 ± 15.3	92.8 ± 3.0	1.4 ± 2.2	98.7 ± 0.2
	Set7	76.4 ± 5.5	50.6 ± 14.4	13.3 ± 22.1	93.4 ± 3.1	2.3 ± 3.7	98.8 ± 0.3
FEDOC vs FEDOC	Set1	64.2 ± 3.6	59.5 ± 3.5	16.7 ± 6.8	90.6 ± 0.9	31.4 ± 7.7	80.2 ± 1.3
	Set2	56.2 ± 2.1	50.5 ± 3.6	10.0 ± 5.8	91.8 ± 2.9	nan ± nan	79.2 ± 1.0
	Set3	59.5 ± 3.4	55.0 ± 4.7	18.2 ± 5.0	90.7 ± 0.6	33.9 ± 7.4	80.5 ± 1.0
	Set4	64.7 ± 3.1	59.5 ± 3.8	17.3 ± 6.4	90.6 ± 0.4	31.9 ± 9.5	80.4 ± 1.2
	Set5	65.8 ± 2.5	63.6 ± 2.6	18.2 ± 6.8	90.6 ± 0.7	33.1 ± 8.4	80.5 ± 1.3
	Set6	58.8 ± 3.0	55.5 ± 3.4	14.2 ± 6.2	90.3 ± 0.6	27.5 ± 8.9	79.7 ± 1.3
	Set7	65.1 ± 2.1	61.0 ± 3.9	20.3 ± 9.9	90.7 ± 0.9	34.9 ± 11.5	81.0 ± 1.9
Impact vs FEDOC	Set1	74.1 ± 6.7	48.3 ± 5.4	10.6 ± 6.1	91.3 ± 3.0	nan ± nan	79.2 ± 0.9
	Set2	65.5 ± 5.3	50.4 ± 4.1	5.5 ± 3.0	92.7 ± 2.7	15.6 ± 9.6	78.5 ± 0.6
	Set3	79.0 ± 6.1	57.2 ± 4.5	13.3 ± 9.6	90.9 ± 1.4	24.7 ± 15.5	79.7 ± 1.7
	Set4	69.9 ± 9.1	48.0 ± 5.6	7.6 ± 4.4	90.8 ± 0.9	17.3 ± 9.2	78.6 ± 0.8
	Set5	78.7 ± 7.9	55.4 ± 7.2	14.2 ± 7.9	92.0 ± 2.8	nan ± nan	80.0 ± 1.3

	Set6	76.0 ± 5.8	56.7 ± 4.5	11.5 ± 7.5	91.6 ± 2.1	25.1 ± 10.7	79.5 ± 1.3
	Set7	76.4 ± 5.5	55.4 ± 5.1	12.7 ± 6.2	91.4 ± 2.5	28.9 ± 11.9	79.6 ± 1.1
FEDOC vs Impact	Set1	64.2 ± 3.6	41.0 ± 16.4	10.0 ± 15.3	95.5 ± 3.6	nan ± nan	98.7 ± 0.2
	Set2	56.2 ± 2.1	44.6 ± 18.0	6.7 ± 13.3	94.2 ± 3.0	nan ± nan	98.7 ± 0.2
	Set3	59.5 ± 3.4	66.1 ± 19.9	20.0 ± 26.7	90.8 ± 1.7	2.5 ± 3.3	98.8 ± 0.4
	Set4	64.7 ± 3.1	41.7 ± 18.0	10.0 ± 15.3	94.6 ± 2.4	1.9 ± 3.1	98.7 ± 0.2
	Set5	65.8 ± 2.5	51.3 ± 24.4	30.0 ± 34.8	93.2 ± 2.8	6.1 ± 6.9	99.0 ± 0.5
	Set6	58.8 ± 3.0	67.5 ± 16.3	23.3 ± 26.0	90.7 ± 1.0	3.4 ± 4.0	98.8 ± 0.4
	Set7	65.1 ± 2.1	46.6 ± 20.6	23.3 ± 26.0	95.9 ± 3.2	5.5 ± 6.5	98.9 ± 0.4

Table S 8 AUCs for stillbirth and neonatal death prediction across all feature sets in FEDOC

		Stillbirth & neonatal death
FEDOC vs FEDOC	Set1	46.6 ± 9.8
	Set2	49.7 ± 8.3
	Set3	46.8 ± 6.8
	Set4	50.6 ± 9.9
	Set5	45.8 ± 7.0
	Set6	46.5 ± 6.9
	Set7	46.0 ± 9.6

Table S 9 Comprehensive performance metrics for C-section prediction across all feature sets

		AUC Train	AUC Test	Sn	Sp	PPV	NPV
Impact vs Impact	Set1	65.3 ± 1.9	65.2 ± 2.9	25.2 ± 5.2	90.1 ± 0.5	54.7 ± 5.7	71.4 ± 1.5
	Set2	57.5 ± 3.2	55.2 ± 3.7	13.3 ± 2.9	90.2 ± 0.2	39.2 ± 4.9	68.3 ± 0.7
	Set3	51.8 ± 2.5	48.7 ± 2.6	3.6 ± 4.7	96.1 ± 4.8	nan ± nan	67.3 ± 0.4
	Set4	65.6 ± 1.9	65.2 ± 3.0	25.6 ± 7.7	90.3 ± 0.5	55.0 ± 6.2	71.6 ± 2.2
	Set5	64.3 ± 2.0	63.4 ± 2.4	20.4 ± 6.6	90.2 ± 0.4	48.8 ± 9.5	70.2 ± 1.8
	Set6	55.9 ± 2.1	52.1 ± 3.3	9.7 ± 2.9	90.6 ± 0.7	32.5 ± 5.5	67.5 ± 0.6
	Set7	65.1 ± 1.5	64.8 ± 2.2	25.2 ± 5.9	90.4 ± 0.4	55.0 ± 6.0	71.4 ± 1.7
FEDOC vs FEDOC	Set1	67.2 ± 3.1	63.0 ± 5.2	25.4 ± 10.3	90.7 ± 0.4	35.8 ± 10.9	84.8 ± 1.8
	Set2	54.7 ± 3.3	49.1 ± 2.7	8.9 ± 4.9	92.1 ± 3.3	nan ± nan	82.2 ± 0.6
	Set3	51.8 ± 2.8	48.6 ± 4.6	3.6 ± 4.5	95.1 ± 4.9	nan ± nan	81.9 ± 0.5
	Set4	66.1 ± 2.8	62.5 ± 5.3	22.5 ± 8.7	90.8 ± 0.5	33.6 ± 9.6	84.3 ± 1.5
	Set5	66.0 ± 4.0	63.0 ± 4.8	21.4 ± 9.8	91.2 ± 0.9	32.0 ± 12.8	84.2 ± 1.6
	Set6	52.3 ± 2.7	46.4 ± 4.8	2.9 ± 4.2	94.6 ± 4.4	nan ± nan	81.7 ± 0.6
	Set7	65.2 ± 3.8	61.4 ± 3.7	22.9 ± 10.4	90.8 ± 1.0	34.2 ± 11.9	84.4 ± 1.9
Impact vs FEDOC	Set1	65.3 ± 1.9	52.3 ± 5.0	15.0 ± 7.5	91.0 ± 1.4	25.2 ± 8.1	83.1 ± 1.1
	Set2	57.5 ± 3.2	50.3 ± 5.0	10.0 ± 3.9	91.1 ± 1.0	19.3 ± 5.9	82.2 ± 0.6
	Set3	51.8 ± 2.5	52.1 ± 4.3	5.0 ± 6.6	95.9 ± 5.0	nan ± nan	82.2 ± 0.5
	Set4	65.6 ± 1.9	50.1 ± 8.1	10.7 ± 6.8	91.2 ± 1.0	19.9 ± 10.9	82.4 ± 1.1
	Set5	64.3 ± 2.0	57.3 ± 5.2	12.9 ± 5.6	91.3 ± 1.6	24.4 ± 9.9	82.8 ± 1.0

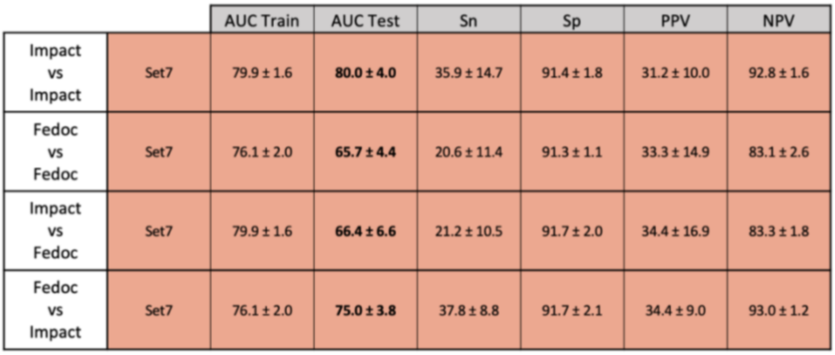
	Set6	55.9 ± 2.1	54.5 ± 2.6	9.6 ± 6.6	91.3 ± 1.3	18.5 ± 11.5	82.2 ± 1.1
	Set7	65.1 ± 1.5	54.1 ± 6.0	14.3 ± 7.7	91.8 ± 1.4	26.2 ± 9.1	83.1 ± 1.2
FEDOC vs Impact	Set1	67.2 ± 3.1	56.8 ± 6.5	14.8 ± 4.1	90.3 ± 0.4	41.8 ± 7.7	68.7 ± 1.1
	Set2	54.7 ± 3.3	53.3 ± 2.9	11.4 ± 7.0	92.4 ± 3.7	nan ± nan	68.4 ± 1.2
	Set3	51.8 ± 2.8	50.2 ± 1.9	5.5 ± 6.0	95.3 ± 4.7	nan ± nan	67.6 ± 0.6
	Set4	66.1 ± 2.8	56.5 ± 3.9	13.4 ± 2.7	90.3 ± 1.3	40.0 ± 4.4	68.3 ± 0.6
	Set5	66.0 ± 4.0	57.4 ± 4.4	14.2 ± 2.5	89.9 ± 0.6	40.3 ± 4.8	68.4 ± 0.6
	Set6	52.3 ± 2.7	52.6 ± 4.5	6.7 ± 6.3	94.2 ± 4.8	nan ± nan	67.6 ± 0.8
	Set7	65.2 ± 3.8	56.6 ± 3.4	13.3 ± 3.7	90.3 ± 0.3	39.1 ± 7.4	68.3 ± 0.9

Discussion of the prediction of alternate outcomes

For low birthweight, IMPACT's results indicate slight prediction improvements when including clinical data and biometrics, rendering the full feature set superior. On the other hand, in FEDOC, Doppler indices are the most predictive, especially when combined with biometrics. In transfer learning contexts, combining Doppler and biometrics is most effective. For preterm births, IMPACT's results had moderate discriminative ability and point to clinical data as most predictive, with improvement when adding Doppler indices. Conversely, FeDoC data yielded poor predictions regardless of feature sets. Transfer learning was poor when predicting in FeDoC and slightly better when predicting in IMPACT. C-section predictions were primarily driven by clinical data, but differences in this feature set among cohorts rendered transfer learning predictions poor. When focusing on preterm births not due to C-sections, the performance was also very poor. Finally, predictions for stillbirth and neonatal death were equivalent to random chance, indicating the limitations of the datasets used in this study

Results considering Gestational Ages of 31-33 weeks

Table S 10 Comprehensive performance metrics for SGA prediction across all feature sets



		AUC Train	AUC Test	Sn	Sp	PPV	NPV
Impact vs Impact	Set7	79.9 ± 1.6	**80.0 ± 4.0**	35.9 ± 14.7	91.4 ± 1.8	31.2 ± 10.0	92.8 ± 1.6
Fedoc vs Fedoc	Set7	76.1 ± 2.0	**65.7 ± 4.4**	20.6 ± 11.4	91.3 ± 1.1	33.3 ± 14.9	83.1 ± 2.6
Impact vs Fedoc	Set7	79.9 ± 1.6	**66.4 ± 6.6**	21.2 ± 10.5	91.7 ± 2.0	34.4 ± 16.9	83.3 ± 1.8
Fedoc vs Impact	Set7	76.1 ± 2.0	**75.0 ± 3.8**	37.8 ± 8.8	91.7 ± 2.1	34.4 ± 9.0	93.0 ± 1.2

Discussion of local scales and Intergrowth-21

As previously reported, Intergrowth-21st standard may fail to detect at-risk SGA infants (birthweight < 10 centile), which may be especially true for western populations comprised by ethnic groups with larger maternal size (59,60). In the case of IMPACT, the proportion of SGA cases using the Intergrowth-21st scale was 10.5% vs. 15.7% when using the local scale. Indeed, the SGA cases detected by both the local standard and Intergrowth-21st are the ones that are at higher risk, whereas those only detected by the local standard are somehow those cases lying in the grey zone.

In Figure S2, we show the comparison of both standards, which reveals a non-linear relation. On the right, we depict the difference between scales assessed at different GAs for individuals with the same weight. The observed differences reinforce the need for caution when applying the Intergrowth-21st tool to populations different from those on which it was trained.

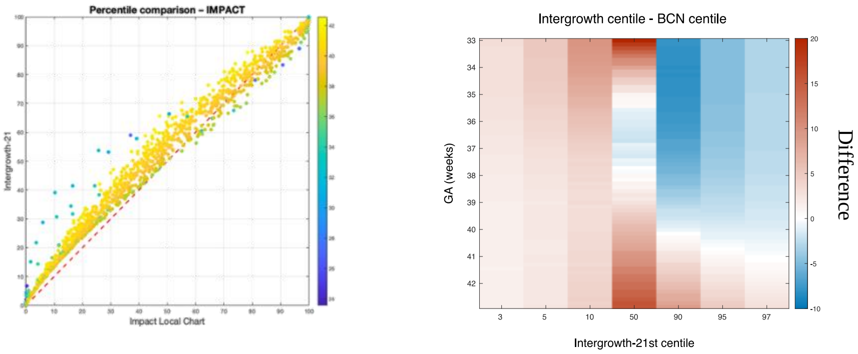


Figure S 2 Comparison of INTERGROWTH-21 and Barcelona's Local Scale

Despite the previously observed differences, and for the sake of comparison and generalization of a model trained on one cohort to the other, in the manuscript we report models' performance in detecting SGA encoded using the Intergrowth-21st scale in both cohorts. Below, a comparison of model's performance across all experiments. As observed, model's performance is superior when using the Intergrowth-21st scale in IMPACT, likely since it's easier for the model to predict these higher-risk cases. In the assessment of model generalizability to the unseen cohort (IMPACT – FeDoC | FeDoC – IMPACT), the models trained at detecting SGA as encoded with the Intergrowth-21st scale in both cohorts systematically outperform those trained at detecting SGA cases encoded using different scales (Intergrowth-21st in FeDoC and Barcelona scale in IMPACT).

Table S 11 AUCs for SGA prediction across all feature sets using different scales

		SGA (Intergr. -BCN st.)	SGA (Intergrowth)
IMPACT – IMPACT	Set1	65.7 ± 4.0	59.4 ± 5.3
	Set2	74.8 ± 4.5	78.7 ± 2.0
	Set3	63.1 ± 4.2	65.8 ± 4.7
	Set4	77.6 ± 2.7	75.0 ± 3.4
	Set5	67.9 ± 5.1	67.4 ± 5.8
	Set6	75.5 ± 5.5	80.3 ± 3.0
	Set7	77.0 ± 5.4	80.3 ± 2.4
IMPACT – FEDOC	Set1	51.0 ± 6.7	56.3 ± 5.9
	Set2	62.9 ± 5.8	67.3 ± 5.5
	Set3	60.9 ± 4.1	61.3 ± 3.8
	Set4	57.4 ± 6.4	64.3 ± 6.1
	Set5	58.8 ± 4.3	63.6 ± 4.6
	Set6	69.3 ± 5.0	70.3 ± 4.5
	Set7	66.3 ± 4.2	71.0 ± 6.5
FEDOC - IMPACT	Set1	54.3 ± 8.0	58.0 ± 6.2
	Set2	66.2 ± 5.4	70.1 ± 4.4
	Set3	59.8 ± 6.4	64.0 ± 4.7
	Set4	62.8 ± 4.8	68.8 ± 4.9
	Set5	59.0 ± 3.5	64.4 ± 3.3
	Set6	69.4 ± 5.4	76.8 ± 3.5
	Set7	63.8 ± 4.3	68.4 ± 6.7

Discussion of baseline characteristics

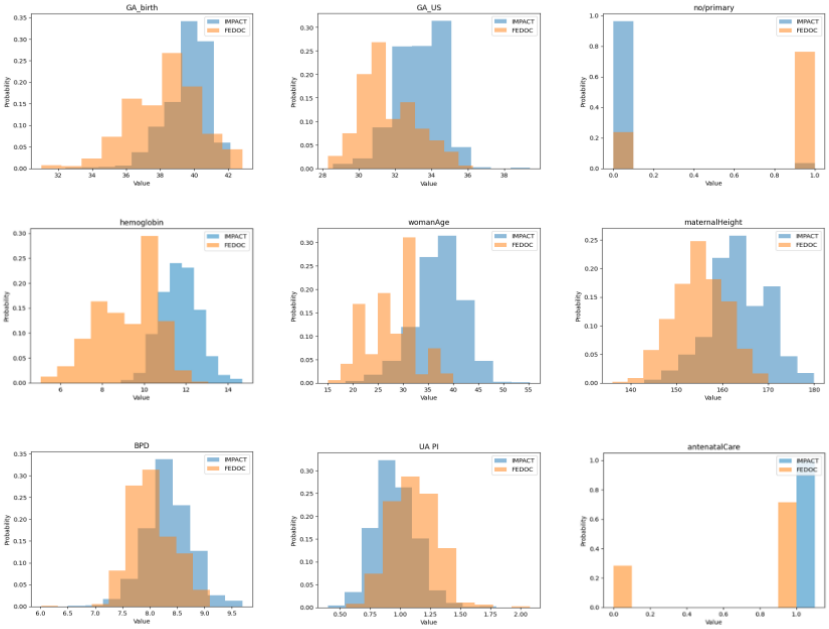


Figure S 3 Visual comparison of baseline characteristics between the IMPACT and FEDOC

We have noted that the cohort from Pakistan is comparatively smaller in size. This is evident across all materno-fetal attributes, as the Pakistani women were recorded as being shorter and lighter in weight. No significant distinction was observed in blood pressure levels. The average length of pregnancies among Pakistani women was also shorter, with the mean gestational age at birth being markedly less than that of the Barcelona cohort.

Clear social disparities were detected in areas such as accessibility to antenatal care, education level, and employment status. Specifically, a third of the Pakistani women lacked access to antenatal care services, their education level was markedly lower, and a majority were unemployed. This contrasts sharply with the highly educated and near-full employment status within the Barcelona cohort.

A higher proportion of women from Barcelona halted risk-associated habits upon confirmation of their pregnancy, whereas approximately half of the Pakistani women continued these habits during pregnancy. Additionally, the Barcelona cohort exhibited a higher representation of nulliparous women, compared to the Pakistani group, where such women were virtually absent.

Hemoglobin levels were markedly lower in the Pakistani group, falling within the range of anemia. This could be due to undernutrition and, given the amount of parous women in this group, suboptimal recovery from a previous pregnancy, among other multiple causes.

3. ASSESSING THE IMPACT OF INTERVENTIONS OF CLUB FOOT INDIA INITIATIVE TRUST ON PATIENTS AND THEIR FAMILIES

Abstract

3.1 Introduction

Clubfoot, a musculoskeletal deformity present at birth, affects a substantial number of infants worldwide. This condition, which causes one or both feet to turn severely inward, necessitates early treatment to restore the child's quality of life. Left untreated, it poses significant risks, including difficulty walking and long-term disability. The exact cause of clubfoot is still unknown, but genetics and environmental factors are considered major contributors.

The standard treatment for clubfoot, known as the Ponseti method, is non-surgical and involves two phases: the corrective and maintenance phase (61). The corrective phase includes plaster casting with weekly changes for 4-6 weeks, followed by an evaluation for a tenotomy. The maintenance phase involves the replacement of the plaster cast with a brace to be worn almost constantly over a three-month span, and then during sleep for up to five years. This cost-effective and non-invasive approach has revolutionized the treatment of clubfoot, especially in low and middle-income countries (LMICs) where access to surgical interventions may be limited (62–64).

However, despite the availability of this effective treatment, access remains a significant barrier in many LMICs. Financial difficulties often prevent families from seeking help, and those who receive care face other long-term obstacles beyond the financial real (65–68). In this context, the role of non-governmental organizations (NGOs) becomes crucial.

One such NGO is CURE International, and specifically, the Cure India Initiative Trust (CIIT), which has been instrumental in addressing these challenges. Started in 2009, CIIT not only provides no-cost treatment for children with clubfoot but also works extensively on other aspects of care. This includes deploying social workers who aid in various areas such as vaccination, water usage, maternal literacy, and more, ultimately aiming to improve outcomes beyond just the clinical perspective (69).

This broader approach aligns with the concept of value-based healthcare, which has gained prominence as it optimizes interventions and focuses on the patient's perspective to identify those interventions that bring the most value throughout the patient's journey. Understanding the impact beyond clinical outcomes is essential, and to this end, monitoring patient-reported outcomes and measurements (PROMs) is key (9). However, it is vital to adapt these PROMs to the specificities of the pathology and the sociocultural context.

In the specific case of clubfoot in India, considering the social stigma that families with affected members may experience, the impact on parents' and children's quality of life, and other social and economic aspects, is especially important (65). These factors are particularly pronounced in families of lower socioeconomic status. A holistic evaluation, integrating a multitude of patient perspectives and extensive questionnaires about patients' psychosocial statuses, can assist in quantifying the social impact of treatments, identifying areas that show pronounced improvement or face challenges.

Yet, there is currently no comprehensive framework that accounts for all the social metrics for measuring the impact an investment has from a social or environmental standpoint. Current frameworks do not measure negative impacts, which are essential for determining the success of an investment. The existing metrics used to measure social impact are incomplete, expensive, lack community involvement, and are ineffective in determining resource allocation. A holistic approach involving community voice and real-time data across various metrics can provide greater insight, ensure high-value healthcare, and make a true community impact.

Institutions such as the International Consortium for Health Outcomes Measurement (ICHOM) have been pioneering in establishing gold standards for these PROMs, with validated questionnaires for various pathologies. The need for such comprehensive evaluation is particularly highlighted by the clubfoot situation in India, and organizations like CURE International and

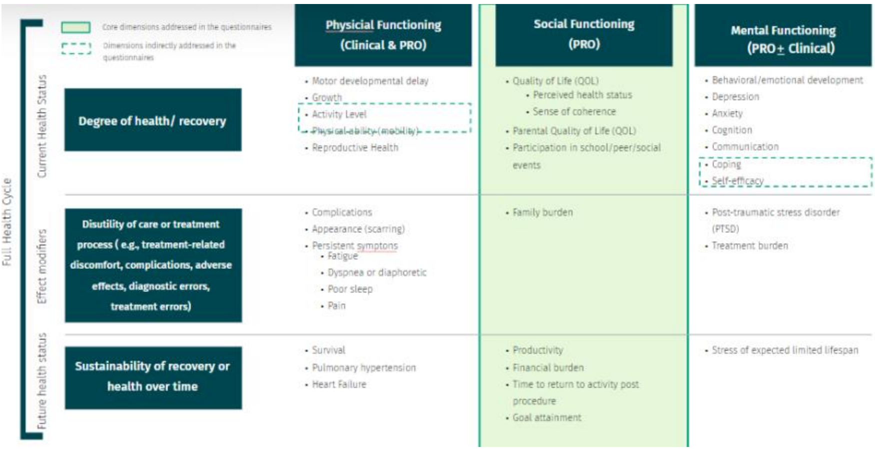
CIIT are leading the way in transforming healthcare delivery by addressing these challenges holistically.

In this paper we present the results and discussion obtained from running a composite of questionnaires to clubfoot and control families in the states of Gujarat and Maharashtra. This corresponds to the baseline characteristics of the cohorts we plan on following up in order to quantitatively and holistically assess the impact of clubfoot treatment.

3.2 Methods

a) Questionnaire selection

In this paper, we employed a multi-faceted methodological approach. Central to this approach was the use of the International Consortium for Health Outcomes Measurement (ICHOM) patient-centered outcomes. As there were no specific questionnaires for clubfoot, we used those for congenital heart disease and congenital upper limb conditions given the similarity (70,71). The perspectives of social workers regarding culturally sensitive considerations further enriched our comprehensive approach to analysis.



🟩 Core dimensions addressed in the questionnaires

▭ Dimensions indirectly addressed in the questionnaires

			Physicial Functioning (Clinical & PRO)	Social Functioning (PRO)	Mental Functioning (PRO± Clinical)
Full Health Cycle	Current Health Status	Degree of health/ recovery	• Motor developmental delay • Growth • Activity Level • ~~Physical ability (mobility)~~ • Reproductive Health	• Quality of Life (QOL) • Perceived health status • Sense of coherence • Parental Quality of Life (QOL) • Participation in school/peer/social events	• Behavioral/emotional development • Depression • Anxiety • Cognition • Communication • Coping • Self-efficacy
	Effect modifiers	Disutility of care or treatment process (e.g., treatment-related discomfort, complications, adverse effects, diagnostic errors, treatment errors)	• Complications • Appearance (scarring) • Persistent symptons • Fatigue • Dyspnea or diaphoretic • Poor sleep • Pain	• Family burden	• Post-traumatic stress disorder (PTSD) • Treatment burden
	Future health status	Sustainability of recovery or health over time	• Survival • Pulmonary hypertension • Heart Failure	• Productivity • Financial burden • Time to return to activity post procedure • Goal attainment	• Stress of expected limited lifespan

Figure 9 Correspondence between dimensions from the ICHOM framework and questionnaires used in this work

The design of our composite of questionnaires was based on various dimensions derived from ICHOM's patient-centered outcomes, as can be seen in Figure 9, as well as social worker preferences and familiarity with the instruments. We also placed an emphasis on the initiative's most significant impacts as reported by social workers and ensured a culturally sensitive adaptation for our target population.

In shaping our study, we aligned the ICHOM domains with validated questionnaires that social workers felt comfortable administering. We emphasized areas such as discrimination and stigma due to their profound influence on patients with clubfoot and their families. Our chosen questionnaires included the WHO-QoL Bref questionnaire (72), the Socioeconomic Status (SES) Udai Pareek revised Scale (73), and the DISC-12 questionnaire for the Discrimination and Stigma Scale (74), adapted for general health issues. We also integrated additional questions from the Lancaster General Hospital questionnaire for assessing Quality of life in children who underwent cardiac surgery to evaluate the financial burden and effects on parents' work productivity due to their child's condition.

We involved social workers in several sessions to foster their understanding of the project and its underlying motivations, ensuring their vested interest in the research. Together, we refined the questionnaire's content, using their insights to shape the final design. This process entailed modifying, eliminating, and adding questions based on their input. Notably, we expanded the list of material possessions in the SES to include two-wheelers and four-wheelers and introduced rented housing as an option. We also adjusted or removed questions pertaining to sexual relations and intimacy.

To ensure translation accuracy and sensitivity, we initially tasked social workers with translating the questionnaire. This was followed by a back-translation to check for any alterations in meaning. Collaborating with the entire team was crucial to ensure the correct interpretation of concepts and prevent any loss of meaning during translation.

For data collection was conducted using Google Forms due to its user-friendly interface, device adaptability, and standardization capabilities. The questionnaire was administered by social workers to families with whom they had established a trustful relationship through repeated clinic visits. This strategy aimed to prevent premature alienation of the families. We chose not to administer the questionnaire to children under three years old, given their limited capacity to contribute meaningfully to our research focus.

Collecting patient-reported outcomes was paramount in our methodology. This approach enriched our comprehensive understanding of the patient experience and added credence to our findings. The resulting data will not only validate our research but also establish a benchmark for future studies, providing a measure for evaluating the effectiveness of new strategies.

During the recruitment phase, we actively sought families from both states who were affiliated with the program. To adjust for potential confounding of environmental variables, the control groups were obtained through proximity matching. This means that the control groups were recruited from neighbors of the same communities as the families with a child suffering from clubfoot. The controls were meticulously selected to match as closely as possible with the target group. This ensured that the comparative analysis remained consistent and reliable.

To gain insight into the progression of families dealing with clubfoot through our program, we needed to overcome the limitation of having data from only a single time point. To tackle this, we grouped the clubfoot families based on the their time enrolled in the program. This resulted in three distinct subgroups: the 'Early' group (0 to 100 days), the 'Middle' group (100 to 500 days), and the 'Late' group (enrolled for more than 500 days).

b) Statistical analysis

Differences between the characteristics of the two groups were evaluated using a T-test or Mann-Whitney U test for continuous variables, depending on the normality of the data. The chi-squared test was used for binary variables.

3.3 Results

In this study, we gathered data from a total of 896 families who completed the entire set of questionnaires, this corresponded to tuples of parent-child responses. Out of these, 404 families were from Maharashtra, from which 210 families were affected by clubfoot, and the remaining 194 were matched control families. In Gujarat, we interviewed 492 participants in total, where 242 families were affected with clubfoot, and the remaining 250 families were our controls. In Table 4, we present the characteristics and scores across various areas of the multiple questionnaires.

Table 4 Characteristics of the whole cohort of study participants

Variables	Control (n=444)	Clubfoot (n=452)	p-value
Demographics			
Parent age (years)	31.58 ± 5.26	31.15 ± 5.28	0.231
Parent sex (female)	248 (55.86%)	169 (37.39%)	p<0.001
Child age (years)	3.19 (1.6 to 5.12)	2.56 (1.1 to 4.16)	p<0.001
Child sex (female)	207 (46.62%)	145 (32.08%)	p<0.001
Age of enrollment (days)	None	56 (19 to 214.5)	
Time in the program (days)	None	763.5 (132.5 to 1262)	
WHO QoL-BREF			
Physical health Parent	16 (13.33 to 16.67)	14 (12 to 16)	p<0.001
Physical health Child	16 (16 to 18.67)	16 (13.33 to 16)	p<0.001
Psychological Parent	15 (12.5 to 16)	13 (11.5 to 14.5)	p<0.001
Psychological Child	15.43 (13.71 to 17.14)	14.29 (12 to 16)	p<0.001
Social relationships Parent	15.67 (14 to 16.67)	14 (12.67 to 16)	p<0.001
Social relationships Child	16 (14.67 to 18.67)	16 (13.33 to 17)	p<0.001
Environment Parent	15 (13 to 16)	13.5 (11.5 to 14.5)	p<0.001
Environment Child	0 (0 to 0.05)	0 (0 to 0.22)	p<0.001

DISC-12			
Unfair treatment Parent	0 (0 to 0)	0 (0 to 0.5)	p<0.001
Unfair treatment Child	0 (0 to 2)	0.5 (0 to 1.5)	0.329
Stopping self Parent	1.33 (0 to 2.75)	1.5 (0.67 to 2)	0.604
Stopping self Child	0 (0 to 0)	0 (0 to 0.18)	p<0.001
Overcoming stigma Parent	0 (0 to 0)	0 (0 to 0)	0.250
Overcoming stigma Child	2 (0 to 2)	1 (0 to 2)	0.011
Positive treatment Parent	1.83 (0.33 to 2.83)	1.67 (1 to 2)	0.700
Positive treatment Child			
Udai Pareek SES scale classes			
Lower class	75 (16.89%)	101 (22.35%)	0.039
Lower-middle class	206 (46.4%)	217 (48.01%)	0.628
Middle class	89 (20.05%)	92 (20.35%)	0.908
Upper-middle class	74 (16.67%)	41 (9.07%)	p<0.001
Upper class	0 (0%)	1 (0.22%)	0.321
Udai Pareek SES scale domains			
Occupation	3 (1 to 5)	1 (1 to 5)	0.016
Education	5 (3 to 6)	4 (2 to 5)	p<0.001
Land	0 (0 to 0)	0 (0 to 1)	0.002
Social Participation	4 (0.9%)	0 (0%)	0.043

Housing	3 (3 to 3)	3 (2 to 3)	p<0.001
Farm Power	1 (1 to 1)	1 (1 to 1)	p<0.001
Material Possessions	6.5 (3 to 18)	6 (3 to 13)	0.047
More than 5 family members	310 (69.82%)	273 (60.4%)	0.003

Table 5 Characteristics of Gujarat participants

Variables	**Control** **(n=250)**	**Clubfoot (n=242)**	**p-value**
Demographics			
Parent age (years)	31.66 ± 5.29	30.45 ± 5.17	0.012
Parent sex (female)	139 (55.6%)	85 (35.12%)	p<0.001
Child age (years)	2.9 (1.58 to 4.77)	2.2 (0.83 to 3.51)	p<0.001
Child sex (female)	98 (39.2%)	80 (33.06%)	0.156
Age of enrollment (days)	None	41 (18 to 131)	
Time in the program (days)	None	544.5 (136 to 1225)	
WHO QoL-BREF			
Physical health Parent	15.43 (13.71 to 16.57)	14.86 (12.86 to 16)	0.001
Physical health Child	16 (14.67 to 16.67)	14 (12 to 15.83)	p<0.001

Psychological Parent	16 (16 to 17.33)	16 (14.67 to 17.33)	0.01
Psychological Child	15 (13 to 16)	13 (12 to 15)	p<0.001
Social relationships Parent	15.6 ± 2.02	14.24 ± 2.22	p<0.001
Social relationships Child	15.33 (14 to 16.67)	14 (12.67 to 15.33)	p<0.001
Environment Parent	16 (16 to 18.67)	16 (14.67 to 17.33)	0.065
Environment Child	14.75 (13.5 to 16)	13.5 (12.5 to 14.5)	p<0.001
DISC-12			
Unfair treatment Parent	0 (0 to 0)	0 (0 to 0.17)	p<0.001
Unfair treatment Child	0 (0 to 0)	0 (0 to 0.5)	p<0.001
Stopping self Parent	0 (0 to 0.75)	0 (0 to 1)	0.172
Stopping self Child	0.5 (0 to 2.62)	1.25 (0 to 2.25)	0.054
Overcoming stigma Parent	0 (0 to 0)	0 (0 to 0.09)	p<0.001
Overcoming stigma Child	0 (0 to 0)	0 (0 to 0)	0.005
Positive treatment Parent	0 (0 to 2)	0 (0 to 1)	0.444
Positive treatment Child	1 (0 to 2.67)	1.83 (0.63 to 2.38)	0.091
Udai Pareek SES scale classes			
Lower class	13 (5.2%)	13 (5.37%)	0.932
Lower-middle class	85 (34%)	111 (45.87%)	0.007
Middle class	78 (31.2%)	77 (31.82%)	0.882
Upper-middle class	74 (29.6%)	40 (16.53%)	0.001

Upper class	0 (0%)	1 (0.41%)	0.309
Udai Pareek SES scale domains			
Occupation	4 (1 to 5)	3 (1 to 5)	0.051
Education	4.5 (3 to 6)	3 (2 to 5)	p<0.001
Land	0 (0 to 1)	0 (0 to 1)	0.060
Social Participation	1 (0.4%)	0 (0%)	0.324
Housing	3 (3 to 3)	3 (2 to 3)	p<0.001
Farm Power	1 (1 to 1)	1 (1 to 2)	p<0.001
Material Possessions	13 (7 to 18)	12 (7 to 18)	0.014
More than 5 family members	166 (66.4%)	140 (57.85%)	0.050

Table 6 Characteristics of Maharashtra participants

Variables	**Control (n=194)**	**Clubfoot (n=210)**	**p-value**
Demographics			
Parent age (years)	31.48 ± 5.24	31.96 ± 5.3	0.366
Parent sex (female)	109 (56.19%)	84 (40%)	p<0.001
Child age (years)	3.71 (1.77 to 5.82)	2.92 (1.62 to 5.25)	0.112
Child sex (female)	109 (56.19%)	65 (30.95%)	0.001

Age of enrollment (days)	None	89.5 (24 to 375)	
Time in the program (days)	None	839 (129 to 1451)	
WHO QoL-BREF			
Physical health Parent	14.57 (12.57 to 17.14)	14.29 (12 to 16.29)	0.150
Physical health Child	15.33 (12.67 to 17.33)	14 (11.33 to 16)	p<0.001
Psychological Parent	16 (13.33 to 18.67)	16 (12 to 16)	p<0.001
Psychological Child	13.79 ± 3.27	12.56 ± 2.79	p<0.001
Social relationships Parent	14.86 (13.71 to 17.14)	14.29 (12 to 16.57)	0.010
Social relationships Child	16 (13.33 to 17.33)	14.67 (12 to 16.67)	p<0.001
Environment Parent	16 (14.33 to 17.67)	16 (12 to 16)	p<0.001
Environment Child	15 (12 to 16)	13 (10.5 to 14.5)	p<0.001
DISC-12			
Unfair treatment Parent	0 (0 to 0.05)	0 (0 to 0.32)	0.007
Unfair treatment Child	0 (0 to 0)	0 (0 to 0.46)	0.032
Stopping self Parent	1 (0 to 2)	1 (0 to 2)	0.060
Stopping self Child	2 (0.63 to 3)	1.67 (1 to 2)	0.166
Overcoming stigma Parent	0 (0 to 0.08)	0 (0 to 0.33)	0.043
Overcoming stigma Child	0 (0 to 0)	0 (0 to 0)	0.873
Positive treatment Parent	2 (0 to 2)	1 (0 to 2)	0.013
Positive treatment Child	2 (0.75 to 3)	1.67 (1.17 to 2)	0.048

Udai Pareek SES scale classes			
Lower class	62 (31.96%)	88 (41.9%)	0.039
Lower-middle class	121 (62.37%)	106 (50.48%)	0.016
Middle class	11 (5.67%)	15 (7.14%)	0.547
Upper-middle class	0 (0%)	1 (0.48%)	0.336
Upper class	0 (0%)	0 (0%)	-
Udai Pareek SES scale domains			
Occupation	1 (1 to 5)	1 (1 to 4)	0.178
Education	5 (4 to 6)	4 (3 to 5)	p<0.001
Land	0 (0 to 0)	0 (0 to 1)	0.007
Social Participation	3 (1.55%)	0 (0%)	0.072
Housing	3 (2 to 4)	3 (2 to 3)	p<0.001
Farm Power	1 (1 to 1)	1 (1 to 1)	0.008
Material Possessions	1 (0 to 3)	3 (0 to 3)	0.329
More than 5 family members	144 (74.23%)	133 (63.33%)	0.022

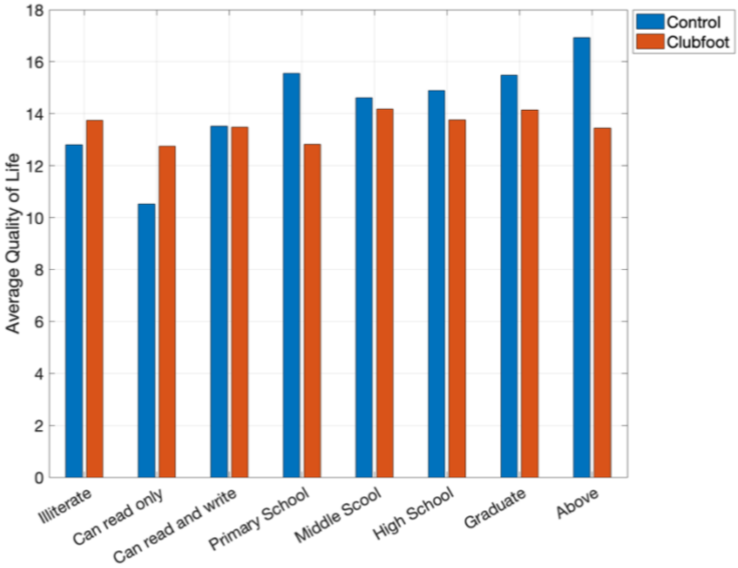


Figure 10 Quality of Life by Maternal Education level

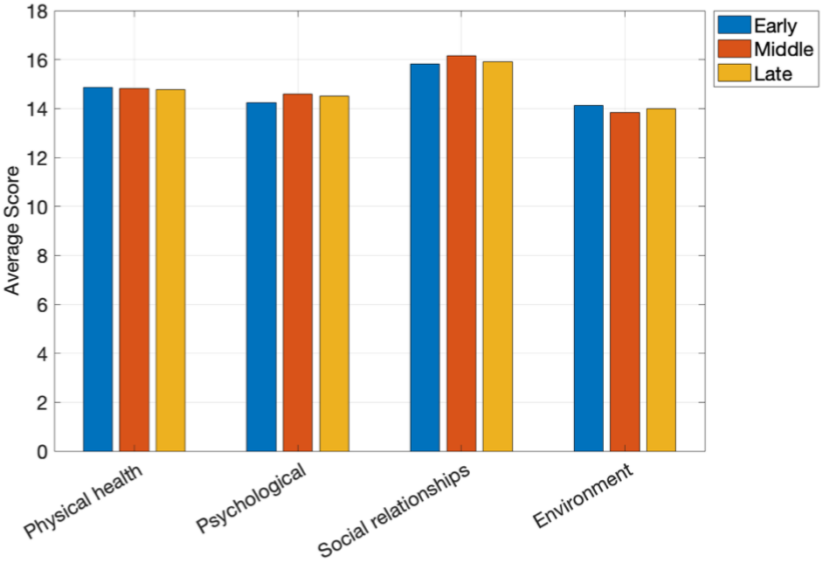


Figure 11 Quality of life of clubfoot affected families by time enrolled in the program

Early corresponds to the period between 0 and 100 days, Middle corresponds to the period over 100 days and below 500 days, and Late corresponds to 500 days and above

We illustrate the distributions corresponding to the scores of the Udai Pareek SES scale in Figure 12. The matching process appears to have been more effective in Maharashtra, as in Gujarat we observe a skew towards lower values for the clubfoot families. Conversely, the control families display a higher SES score.

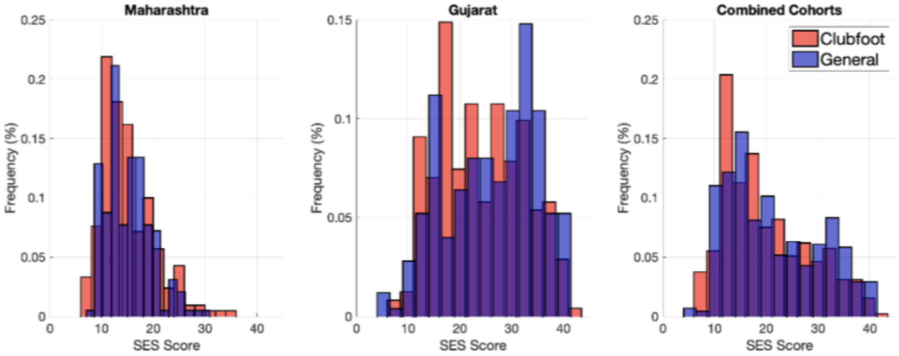
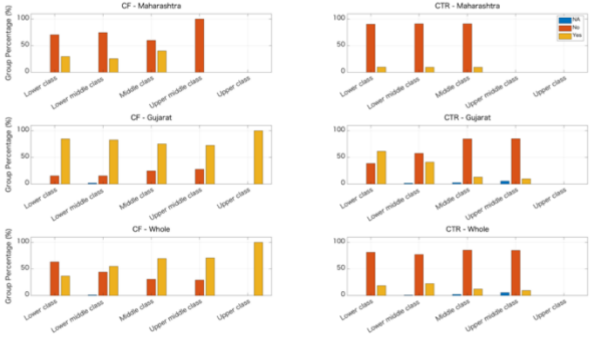


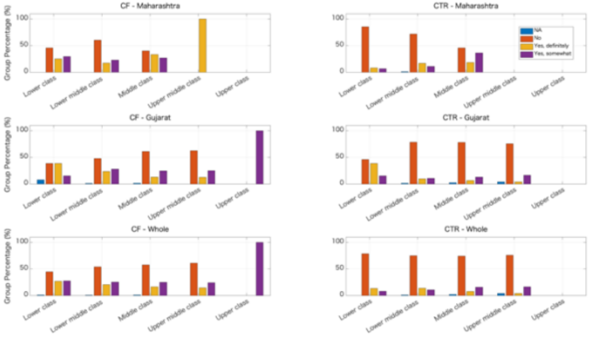
Figure 12 Udai Pareek SES scores

We also acknowledge that the domain of Overcoming stigma in DISC-12 was extremely sparse, containing the largest amount of non-answered questions out of every area. Making it hard to make assessments.

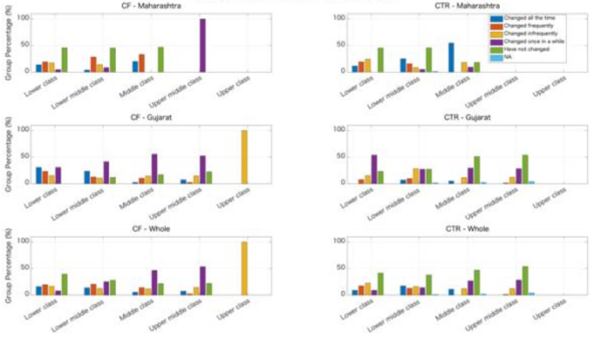
In the past 12 months, has your child had emotional, developmental, or behavioral problems for which he/she received treatment or counseling?



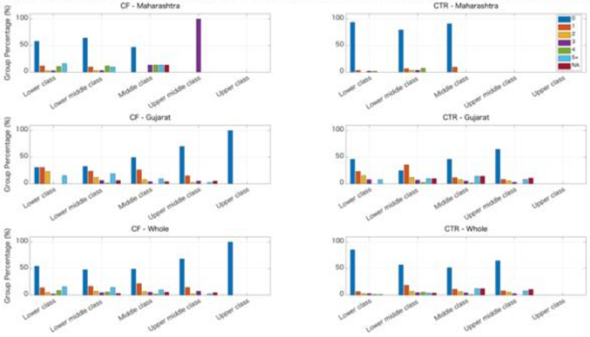
In the past 12 months, has your child's health limited or prevented him/her in any way to do things that most children of the same age can do?



In the past 12 months, how often have your child's health care needs changed?



How many children under the age of 18 (not including the child that you answered the questions about in this survey) have special health care needs that require them to see multiple health care providers?



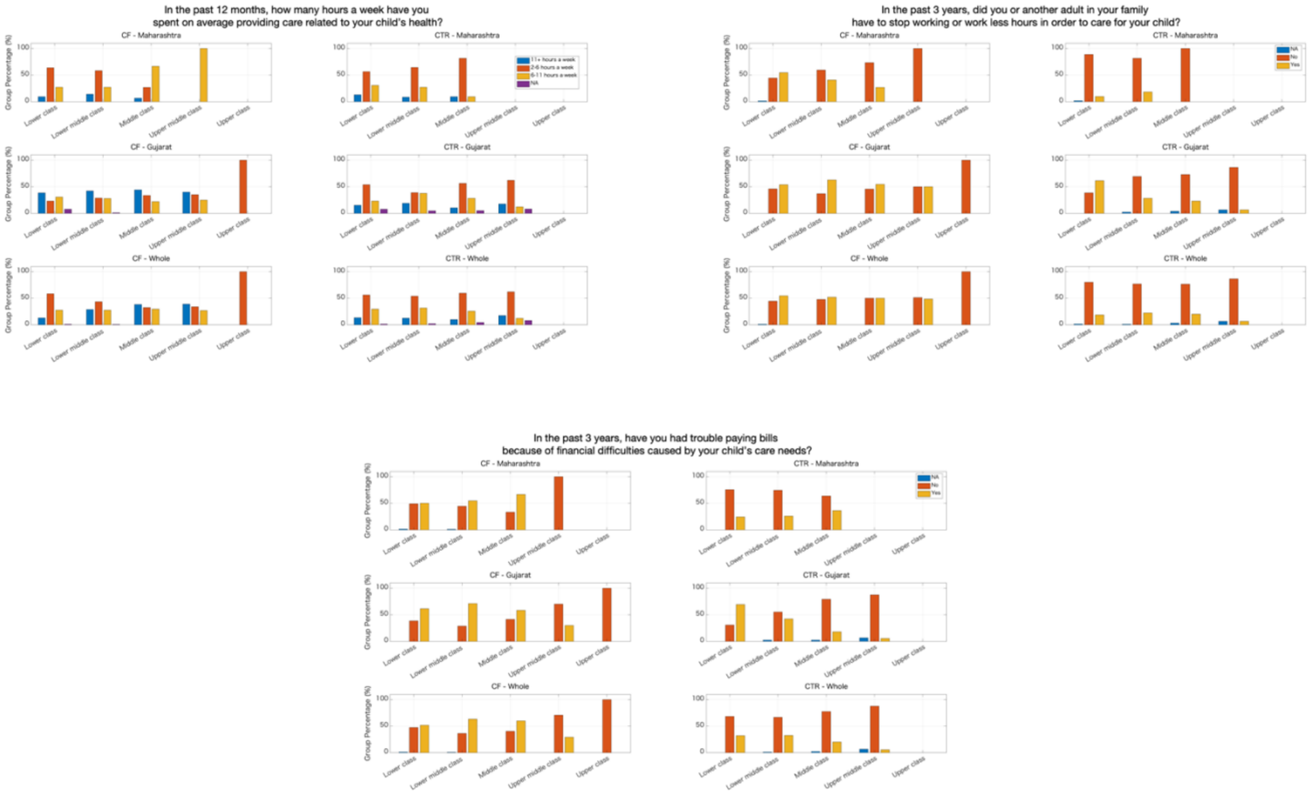
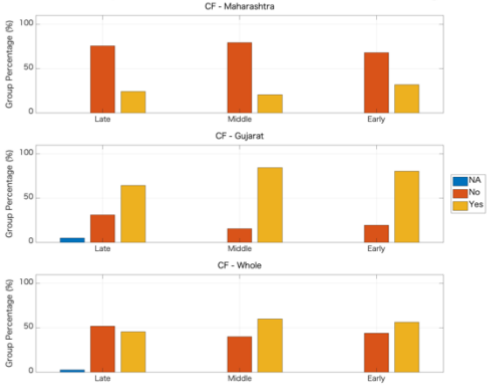
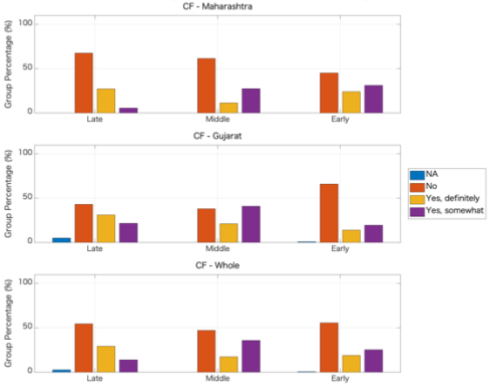


Figure 13 Responses to the LGH Questionnaire stratified by Socio-economic class

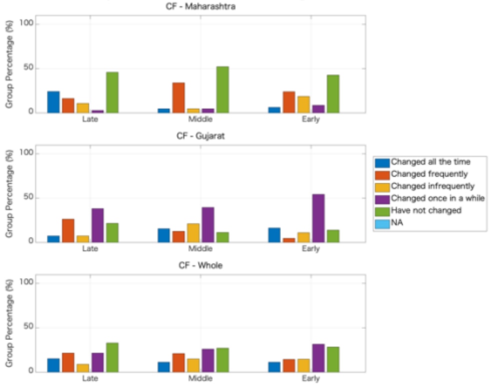
In the past 12 months, has your child had emotional, developmental, or behavioral problems for which he/she received treatment or counseling?



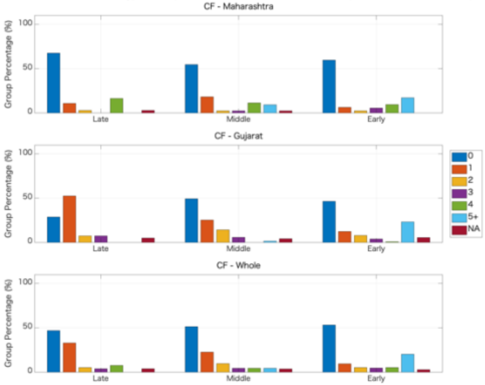
In the past 12 months, has your child's health limited or prevented him/her in any way to do things that most children of the same age can do?



In the past 12 months, how often have your child's health care needs changed?



How many children under the age of 18 (not including the child that estions about in this survey) have special health care needs that require them to see multiple l



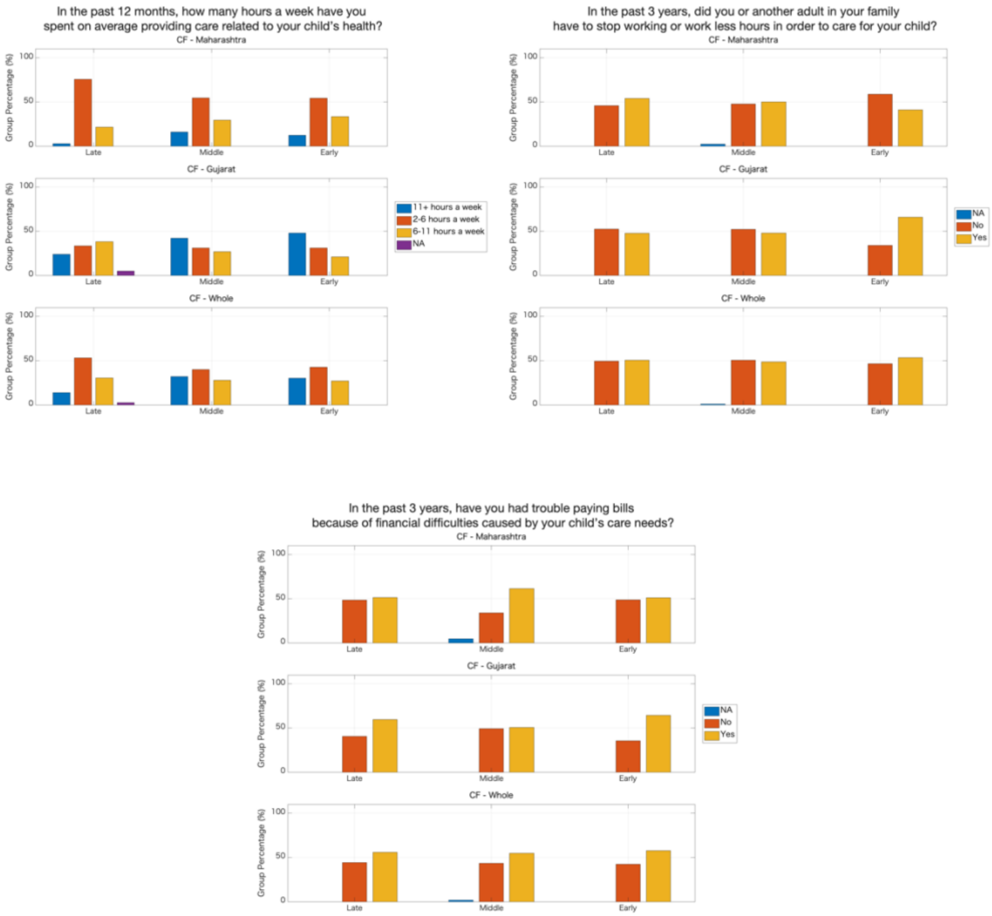


Figure 14 Responses to the LGH Questionnaire by time enrolled in the program

3.4 Discussion

To begin with, regarding the differences between states, we observed that Gujarat families tend to have lower levels of education (Table 5), this demographic potentially representing a group with high vulnerability. In contrast, the state of Maharashtra exhibits higher levels of education but suffers from a lower overall socioeconomic status (Table 6).

Moreover, statistical analysis revealed significant differences in the Quality of Life (QoL) of children with clubfoot and their parents when compared to controls across all dimensions of the questionnaire (Table 4). These differences were particularly marked in the child's environment score in Gujarat (Table 5), in the parents' physical health and in psychological scores for both parents and children in both states (Table 5, Table 6). The most profound differences were observed in the psychological domain, implying the adverse psychological impact of clubfoot on affected families.

The data reveals a direct correlation between maternal education and QoL (Figure 10), with lower education levels generally observed in families with a clubfoot child. The correlation is especially pronounced within the psychological domain, reinforcing the idea of the significant mental health burden associated with having a child with clubfoot.

Despite the significant challenges these families face, there was a noticeable sense of positive discrimination towards families with a clubfoot child in both states (Table 4). This might be tied to the gratitude these families feel towards the program and the benefits they derive from it. In both Maharashtra and Gujarat, higher scores were noted in the domain concerning overcoming stigma, with a higher difference in Maharashtra, although these differences were not statistically significant (Table 5, Table 6, and supplementary Figures 4 and 5).

The supplementary questions provided further insight into the factors contributing to the lower QoL in clubfoot families (Supplementary Figure 6). Approximately 55% of children with clubfoot required additional treatments or counselling for emotional, behavioral, and/or developmental problems, in stark contrast to only about 18% of the control group. Furthermore, around 45% of these children had limitations due to their health and more changing medical needs compared to controls.

An additional strain on these families is evident in the statistic that more than 50% of the families with a clubfoot child have more than one child with special health needs. Some families have as many as five dependants with various health problems. This situation has had a tangible impact on the family structure and economics, with more than 50% of the families with a clubfoot child reporting that an adult had to stop working or reduce their work hours to care for the child with special health needs. The demands of care are significant, with over half of these families spending more than six hours a week on health-related care for their child and 30% spending more than 11 hours weekly. This increased time commitment to care further compounds the financial difficulties faced by these families, with around 60% of them experiencing difficulty paying bills due to their child's healthcare needs (Supplementary Figure 6).

A deeper analysis of these results, taking into account the socioeconomic status (SES) of the clubfoot families, further illuminates the issue (Figure 13). Children from middle and upper-middle-class families were more likely to receive treatment than those from lower SES families, perhaps due to lower access to healthcare. Over 70% of children from higher SES families received counselling or treatment for emotional, behavioral, and/or developmental problems. This percentage is almost double that of children from lower SES backgrounds. Additionally, families with lower SES tended to have more children with special healthcare needs, and a higher percentage had to stop working or reduce their work hours to accommodate their child's needs. Interestingly, lower middle and middle-class families reported the most financial difficulties due to their children's healthcare issues (Figure 13).

When we analyze these findings by state, the results vary (Figure 13). In Gujarat, the influence of SES on the clubfoot families appears more pronounced. There is a higher percentage of children with additional healthcare needs in Gujarat, and a higher perception of the child’s disability level compared to other children of their age in lower SES families. While families in Maharashtra tend to allocate more healthcare intervention time in middle and upper-middle classes, families from lower SES in Gujarat spend more time providing care. Additionally, a higher percentage of adults in Gujarat had to reduce their work hours or stop working entirely due to their child's health needs, irrespective of SES. However, in Maharashtra, this trend was most evident in lower SES families. In general, families in Gujarat faced greater difficulties in paying bills due to their children's healthcare needs.

When we factor in the duration of a family's enrolment in the program, categorized as early, middle, late, we find an interesting correlation (Figure 14). Families in the early stages of the program, and those in the late stages, reported lower perceived disability, likely due to the children's younger age and the intervention's beneficial effects, respectively. Long-term enrolment in the program was also associated with a reduction in the time spent providing care for their children’s needs and a decrease in the need for treatment and counselling for developmental, emotional, and behavioral conditions. Despite these improvements, there wasn't a similar reduction in the financial burdens such as job adjustments or difficulties paying bills. However, the combined effects of less time dedicated to childcare and decreased perceived disability indicate the positive impact of CURE India’s intervention on the families' overall wellbeing.

3.5 Limitations and Future work

The limitations of this study stem from several factors. Primarily, the social vulnerability of the families involved introduces a response dependency when gauging the perception of the child's care. Additionally, the availability of only cross-sectional data limits the measurement of the patient's evolution due to time constraints with data collection. Despite this, participants can be segmented into different stages of the health cycle.

Further complications arise from the questionnaires, which were limited due to the counselors' familiarity with the questions. To facilitate counselors' focus on social impact factors, the questions were modified. However, these changes necessitated adaptation to contextual and cultural cues for the population to which the questionnaires were administered.

The role of the counselors is crucial in this context. Their expertise is required to build trust and rapport with families, ensuring both the accuracy of responses and the families' return for continued treatment.

Another limitation lies in the timing of questionnaire administration. These were given after successful intervention, creating a dearth of ground truth when comparing to unsuccessful interventions. Furthermore, there was no proper baseline set, with a specific period from enrollment to questionnaire administration, as would typically occur in a prospective trial.

3.6 Conclusions

The paradigm of value-based healthcare (VBHC) emphasizes patient-centered outcomes, advocating for a broader, more encompassing perspective that includes patient-reported outcomes for a more sustainable, high-value care approach. This principle is intertwined with the importance of recognizing local contextual challenges, where geographical differences within the same country can yield varied results in healthcare programs, necessitating tailored solutions to enhance accessibility and adherence.

Meanwhile, the sustainability of health outcomes is underpinned by ongoing patient participation in these programs, and while a single time point may provide promising results, continuity in assessing outcome domains throughout the health cycle is pivotal. This comes to the fore in the efficacy of interventions, especially amongst vulnerable segments. Interventions have shown tangible improvements in the quality of life of the patients and their families. Nevertheless, disparities persist between the patient population and control groups, and while there is a clear trend towards improved outcomes over time, the financial burden on families, particularly those in lower socioeconomic and educational brackets, remains a significant challenge.

Moreover, interventions have elicited a positive treatment bias, with clubfoot patients experiencing better outcomes in positive discrimination dimensions due to the extended support from social workers. Their roles extend beyond mere treatment, helping families navigate the healthcare system and educating them about vital health practices like vaccinations and safe water usage.

Cultural considerations in the development of questionnaires are crucial, requiring real-world adaptation and buy-in from communities and counselors to ensure the appropriate implementation of the study. In essence, this multifaceted approach, merging patient-centered VBHC, adaptable learning, sustainable health outcome tracking, culturally conscious questionnaire design, and targeted support for vulnerable segments, promises a more effective, inclusive, and equitable healthcare landscape in our future follow-up studies.

Considering these findings, it is crucial to acknowledge the importance of measuring beyond just clinical outcomes in Low- and Middle-Income Countries (LMIC). While clinical markers are key to assessing health progress, they do not capture the entirety of a patient's wellbeing or the socio-economic impact of health interventions. By including factors such as financial stability, social support, and education, we get a more holistic view of health and can thus better allocate resources for optimal benefit. This approach promotes high-value care by identifying high-impact investments, pinpointing risk populations, and discovering potential areas for intervention.

This expanded measurement strategy also fosters collaborative learning between different regions and communities. By sharing success stories and learning from challenges, regions can borrow and adapt effective practices to their local context. This kind of collaboration is pivotal in crafting tailor-made strategies that truly respond to the needs and realities of specific populations.

Furthermore, it is instrumental in offering a clear measure of impact for the vulnerable populations caught between clinical outcomes. For instance, though clinical improvement might be noted, it is necessary to identify if the patient still suffers due to ancillary burdens such as transportation costs, time off work, or stigma from the disease. A comprehensive analysis such as this is fundamental to illustrate the full picture of the patient's journey, thereby leading to better-targeted and more impactful interventions.

Ultimately, this broader perspective of health outcomes is not just necessary for better patient care; it is also crucial for funding and sustainability. It enables NGOs to demonstrate a wider societal impact of their interventions, moving beyond solely medical gains. This broader societal value proposition could be more appealing to potential funders, who are increasingly interested in supporting projects that deliver comprehensive benefits to communities. Consequently, such an analytical approach would contribute to securing financial resources that sustain NGOs' activities, leading to healthier communities in the long run.

3.7 Supplementary Material

a) DISC-12 score distribution

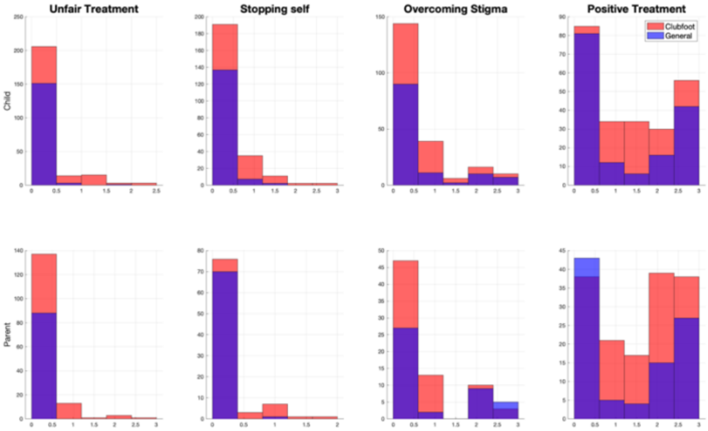


Figure S 4 Gujarat DISC 12 distributions for each of the four domains

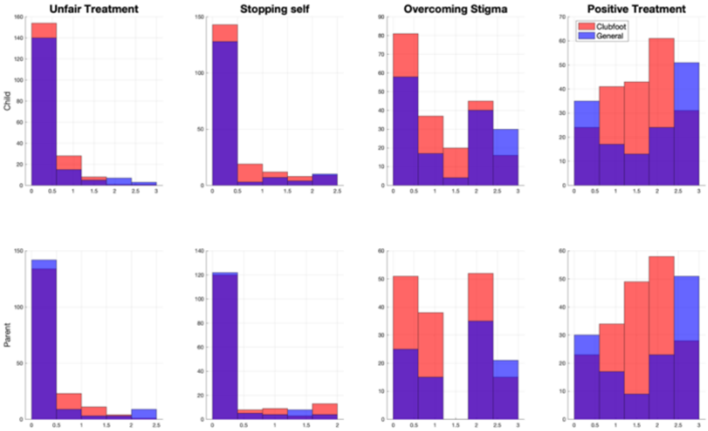
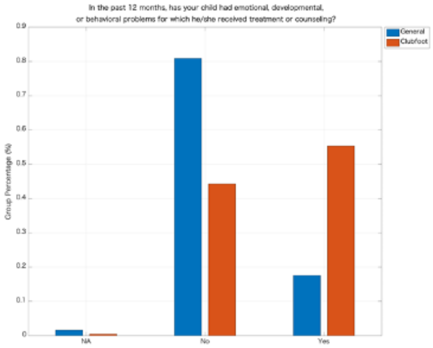
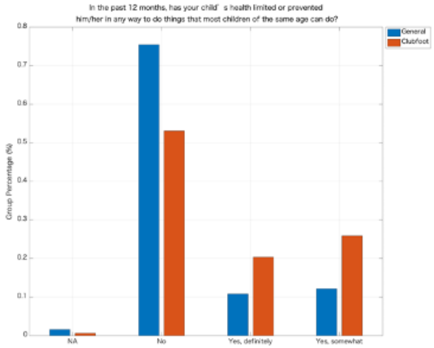


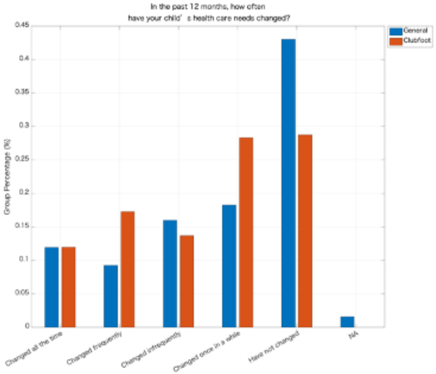
Figure S 5 Maharashtra DISC 12 distributions for each of the four domai



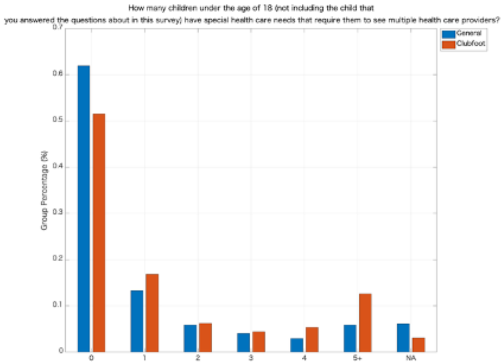
In the past 12 months, has your child had emotional, developmental, or behavioral problems for which he/she received treatment or counseling?



In the past 12 months, has your child' s health limited or prevented him/her in any way to do things that most children of the same age can do?



In the past 12 months, how often have your child' s health care needs changed?



How many children under the age of 18 (not including the child that you answered the questions about in this survey) have special health care needs that require them to see multiple health care providers?

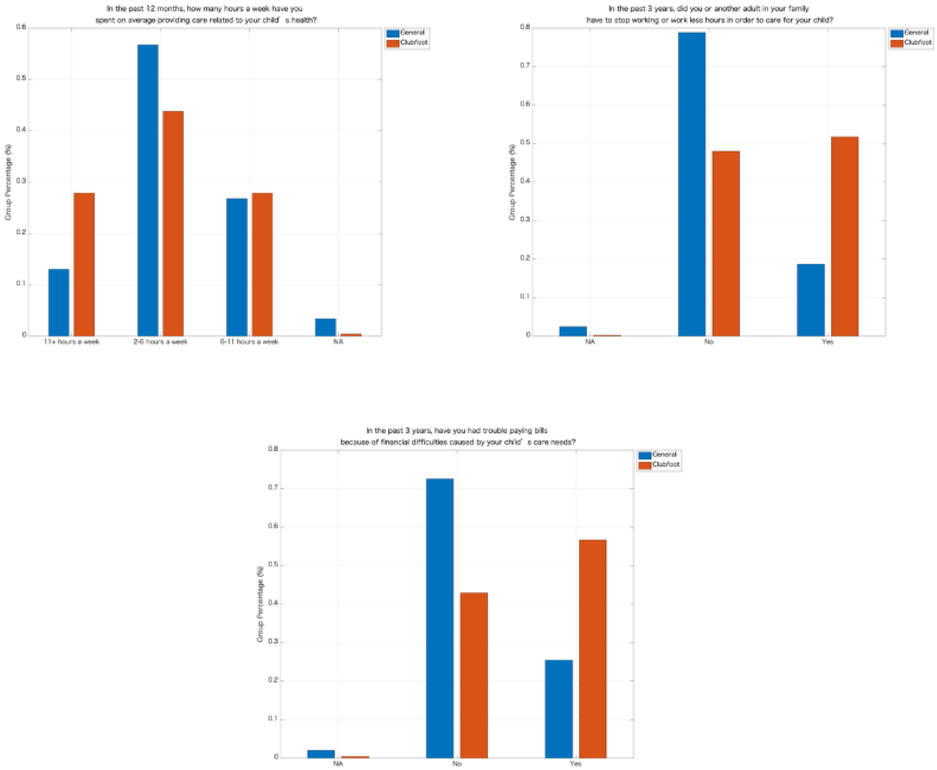


Figure S 6 Responses to the LGH Questionnaire by Clubfoot or Control

4. INCREMENTAL MULTIPLE KERNEL LEARNING

Abstract

Unsupervised learning algorithms are becoming increasingly important in the big data era due to their ability to extract knowledge from large amounts of unlabeled data. Most existing unsupervised learning algorithms that allow for data fusion using kernel methods suffer from slow learning speeds due to high computational costs. To overcome this limitation, we propose a novel unsupervised learning method called incremental Multiple Kernel Learning for Dimensionality Reduction (iMKL-DR). This algorithm builds upon unsupervised MKL, and thus can also fuse information from multiple sources with different dimensionality. Our proposed method can learn the underlying data representation from heterogeneous sources without presupposing any prior knowledge about the data. Experimental results show that our approach can obtain embeddings on par with its batch alternative while reporting smaller computation times.

4.1 Introduction

Multi-modal representation learning is growing in importance. Many algorithms are focused on the problem of integrating heterogeneous data to obtain a unified representation that can leverage the information captured by the different modalities of acquisition. This is particularly crucial in healthcare or biomedical context, where data collected from a single individual presents an intrinsic heterogeneity due to the source itself (imaging, biomarkers, cell counts, gene expressions, demographics, etc.), making it difficult to unify all these descriptors in a simpler representation.

There are many types of algorithms that aim at tackling the problem of representation learning with heterogeneous data, but we focus on the field of kernel methods, as they provide an interesting trade-off between performance and dataset size. In comparison, deep learning architectures, albeit being the most common in practice as they can capture extremely complex patterns, they require vast amounts of observations to train. These dataset sizes might not be available in certain biomedical settings, as healthcare data is expensive to obtain.

Kernel methods leverage the use of a kernel matrix to represent the pairwise similarities between observations, using a kernel function. This approach holds the advantage of mapping the original high-dimensional input data into a new, potentially infinite-dimensional, feature space. This transformed space, often referred to as a 'kernel space,' is designed to make the data more separable, making complex patterns more visible. In essence, kernel methods aim to learn a kernel matrix that captures the similarities between instances, such as patients in a medical context, as accurately as possible compared to the original input data. From this, they then derive a lower-dimensional representation that emphasizes the most significant characteristics of the data. In equation 1, we illustrate how a kernel matrix can be computed using one of the most common kernel functions, known as radial basis or Gaussian.

Equation 1

$$\boldsymbol{K}_{\boldsymbol{m}\left(x_{\{m,i\}},x_{\{m,j\}}\right)} = \mathbf{exp}\left(-\frac{\left|\left|\boldsymbol{x}_{\{\boldsymbol{m},\boldsymbol{i}\}} - \boldsymbol{x}_{\{\boldsymbol{m},\boldsymbol{j}\}}\right|\right|^{2}}{2\ \sigma_{m}^{2}} \right)$$

An already established algorithm in this scenario is multiple kernel learning, for which many formulations exist. This holds true especially in the supervised case, where formulations allow for the training with a vast number of kernels and instances (75–77).

In this chapter we aim at improving the formulation used in Chapter 1, by Lin et al (27), which as all batch unsupervised MKL formulations suffer from poor scalability and are thus unable to train with very big datasets.. In a nutshell, this occurs due to the quadratically increase in computational cost due to the nature of the kernel matrix which represents each observation of a dataset as a row and column of a square matrix.

Nonetheless, we acknowledge two implementations that aim at resolving this issue, namely, TUMK-ELM and fMKL-DR (78,79). TUMK-ELM is capable of learning at very high speeds, but it presupposes clusters within the input data, which might hinder the detection of transitions that can occur during disease worsening from normalcy. On the other hand, fMKL-DR, which builds on top of MKL-DR, implements matrix chain multiplication ordering to achieve a slightly faster convergence to the solution, but the limitation when dealing with a very large number of data observations remains.

In this last chapter, we want to add an additional tool to the unsupervised repertoire that does not have the shortcomings of current implementations. For this reason, we propose a rather simple formulation that builds on top of previous work by Li and colleagues (80). Our proposed solution addresses the scalability issue by innovatively applying an iterative retraining method to unsupervised Multiple Kernel Learning (MKL). The central idea here is to incrementally introduce small data batches at each iteration, which enables MKL to handle larger datasets effectively. This iterative, incremental process also allows the model to learn from new datapoints and adapt to progressively larger data volumes, where previously a full batch retraining would have been necessary.

We named our approach incremental Multiple Kernel Learning for Dimensionality Reduction (iMKL-DR in short). The key contributions from our work are:

- **Heterogeneous data fusion**. Profiting from data kernelization, our algorithm can merge data with different dimensionality, and typology without any label information.

- **Scalability**. Our approach presents a reduction in computational costs, potentially allowing for training with more instances.
- **Non-presumptive dimensionality reduction.** The spectrum capturing disease progression, which ranges from normalcy to disease, is not a discrete process but a grey area between both (25). This makes diagnosis of complex syndromes challenging, since clinicians need to integrate plenty of data from a single individual and compare it to similar ones. This renders algorithms that assume separated clusters in the data (like TUMK-ELM) suboptimal in these cases. In contrast, iMKL-DR could improve the clinical decision process by presenting the clinicians with an agnostic, continuous and yet informative representation of the individual status in comparison to others.

We compared the resulting embedding quality in both a benchmark dataset and a clinical dataset. First, a hypertrophic cardiomyopathy dataset (26), second, a dataset of manuscript digits from which multimodal features have been extracted (81).

We organized this chapter as follows: First, in the methods section, we describe in detail the formulation on which iMKL-DR is based and an in-depth data description and preparation. Next, in the results section, we provide the results obtained in the datasets mentioned, elaborate on the insights provided by our approach, and compare them with classical MKL-DR. Lastly, we present the discussion, conclusions and future work.

4.2 Methods

a) iMKL-DR Formulation

This section describes the steps of our proposed iMKL-DR algorithm. The process starts by setting up the "seed" space, followed by running a batch MKL algorithm, and ends with the incremental addition of new observations to the learnt space. The key principle of iMKL-DR lies in iteratively integrating new information from new samples into the seed space, leading to a model that can be dynamically updated and does not require a full batch retraining when a new set of observations is acquired. We provide an overview in Algorithm 1 and Figure 15.

Algorithm 1: Incremental Multiple Kernel Learning (iMKL-DR)

Input: Kernel Matrices for the seed and growth data

Output: Space of reduced dimensions containing the previous and new samples

1 **Initialization:** Select a set of initial data to form the "seed" space. Kernelize the multiple features and input them to the batch MKL algorithm. Set bandwidths for the kernel computation and global affinity matrix to the square root of the number of observations.

2 **Batch MKL-DR:** Process the kernelized data to produce the reduced-dimension output space, projection matrix **A**, Similarity Matrix **SWB**, and feature weights vector **betas**.

3 **Projection:** Project new observations to the seed space using the previously obtained **A** matrix and **betas** vector.

4 **Neighborhood Detection:** Identify the samples from the seed space whose neighborhood has been modified by the new samples. The number of nearest neighbors (NNs) for each observation in the seed space is a parameter that can be adjusted.

5 **SWB Update:** Update the Similarity Matrix **SWB** based on the changed neighborhoods, following the method proposed by Li and colleagues.

6 **Betas Update:** Update **betas** using the second optimization procedure in batch MKL-DR.

7 **Final Output Space and Projection Matrix Calculation:** Use the updated **SWB** and **betas** to produce a final projection matrix **A** and output space that incorporate the new samples and their information.

8 **Iteration:** Repeat Steps 3 to 7 for each batch of new observations.

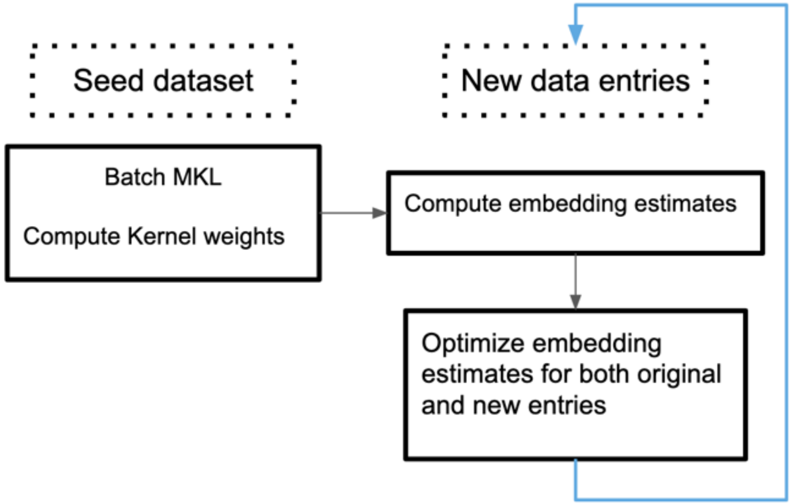

Seed dataset
New data entries
Batch MKL
Compute Kernel weights
Compute embedding estimates
Optimize embedding estimates for both original and new entries

Figure 15 iMKL overview

I. Pre-processing and Seed Space Initialization

The first step in our approach involves selecting a set of initial data that will generate the "seed" space. This seed space will later be expanded upon with new observations. The multiple features of this seed data are kernelized and fed into the batch MKL algorithm.

The primary parameters to be determined are the number of nearest neighbors that regularize the variance for the feature kernels and the global affinity matrix. Empirically, assigning them to the square root of the number of observations has yielded satisfactory results, as per existing literature (16). Running the batch MKL block using the kernels obtained from the seed data then yields an output space of reduced dimensions, a projection matrix **A**, a Similarity Matrix **SWB**, and the weight vector **betas** for different features.

To provide a brief description of the batch MKL algorithm, once all features are kernelized, the goal is to optimize a projection matrix **A** and a feature weight vector **betas** using a two-step optimization strategy. This first step can be solved by a generalized eigenvalue problem framing it as a trace ratio problem, which yields an explicit solution using two auxiliary similarity matrices **SWA** and **SWB**.

The newly obtained similarity Matrix **SWB** serves a crucial role, as we make a correspondence to the matrix **Q** in the incremental algorithm described by Housen Li et al. We provide a further explanation of Li’s algorithm in section III.

After obtaining the projection matrix **A**, a convex optimization problem is set to solve the feature weight values to be stored in the vector **betas**. The batch algorithm would now be finished as observations can be mapped to a low dimensional space using the obtained **A** and **betas**.

II. Projection

New observations are first added to the model by projecting them. This projection of new data points to the seed space is done using the previously derived projection matrix **A** and feature weights, as shown in equation 2.

Equation 2

$$F = \sum_{m} \beta_m K_m \cdot A$$

At this stage, no new information from the new samples has been incorporated into the model; they have simply been positioned within the existing seed space.

III. Neighborhood Detection and Matrix Update

The algorithm then proceeds to assess the impact of the newly introduced observations. This is accomplished by identifying the samples in the seed space whose neighborhoods have been modified by the new observations. The number of nearest neighbors (NNs) for each observation in the seed space is an adjustable parameter.

The next step involves updating the Similarity Matrix **SWB** based on the detected changes in neighborhoods using the method proposed by Li et al (80).

Li’s algorithm is composed of two sub-steps that we explain as follows. First, to allow for the identification of those samples whose neighborhoods have been changed by newly coming ones, a pairwise similarity matrix is obtained using a distance metric between the seed and incoming samples. This metric is generalizable to different spectral embedding formulations. In our case, we define it as the Euclidean distance between points in the embedding once we project the newly coming samples. Secondly, once these neighborhoods have been defined, we can refine the predicted coordinates using an orthogonal iteration method on the **SWB** matrix obtained from batch MKL.

IV. Betas Update and Final Output

Once the **SWB** matrix is updated, the weight vector **betas** is updated through the second optimization procedure in batch MKL. With the updated **SWB** and **betas**, a final **A** matrix and output space are computed. These new elements incorporate the information from the new samples.

V. Iteration

This process is iterative, and the cycle repeats, adding new samples each time. The result from one iteration can be used as the seed for the next cycle, allowing for continual updating and improvement of the model.

b) Data description and preparation

Regarding dataset selection, we used the Multimodal Digits dataset to have an open and easily accessible benchmark dataset for peers to replicate our results, and the Hypertensive dataset to have a medium sized and easily separable dataset, with complex clinical descriptors.

I. Multimodal Digits

This dataset consists of 2000 observations of digits from "0" to "9" (200 observations for each digit) extracted from Dutch utility maps. There are six feature modalities for each digit:

- 76 Fourier coefficients of the character shapes.
- 216 profile correlations.
- 64 Karhunen-Love coefficients.
- 240-pixel averages in 2 x 3 windows;
- 47 Zernike moments.
- 6 morphological features.

When analyzing this dataset, we computed a kernel matrix for each feature type using a radial basis function. This dataset is open and free to use by others, it can be accessed at https://archive.ics.uci.edu/ml/datasets/Multiple+Features.

II. Hypertensive dataset

This dataset resulted from echocardiography studies from 189 clinically managed patients with hypertension and 97 healthy individuals without hypertension. Strain traces of the left ventricle and atrium were obtained using speckle-tracking analysis. Aortic and mitral blood pool pulsed-wave Doppler and mitral annular tissue pulsed-wave Doppler velocity profiles were obtained. These whole–cardiac cycle deformation and velocity curves were used as input, resulting in 11 highly dimensional features.

c) Experiments and evaluation metrics

To assess the performance of iMKL-DR, we launched the same two experiments with each dataset. First, to test whether new observations were correctly positioned in the grown space, we trained a seed space with a subset of each class and grew the space with the remaining ones. Secondly, to assess the performance when adding unseen classes, we trained on a subset of classes and added the rest.

Regarding the parameter settings, we established the number of observations that will be added at each growth iteration to 15 for the digits dataset and to 5 for the hypertensive dataset. Regarding the neighborhood sizes when detecting incoming samples, we set it to 15 neighbors for the digits dataset and to 5 for the hypertensive dataset. We've also conducted a comprehensive exploration of the parameter space for the digits dataset, as detailed in the results' Section a)III. This extensive analysis includes evaluating the sensitivity of our model to varying parameters. Specifically, we have examined the influence of the number of neighbors considered when projecting new patient data, along with the effect of alterations in the size of the seed space.

Evaluating the embeddings produced by unsupervised manifold learning approaches can be difficult, as there are no ground truth labels. In our setting, this problem gets further compounded by the multimodal nature of the data, making it hard to naively implement most of the current metrics. To keep the evaluation of our technique clear, we used a rather simple methodology by visualizing the embeddings to make qualitative assessments of grouping of classes in different dimensions.

4.3 Results

a) Multimodal Digits

To allow for comparisons, we would like to first add a visualization of the space obtained from batch MKL using all observations. This is illustrated in Figure 16.

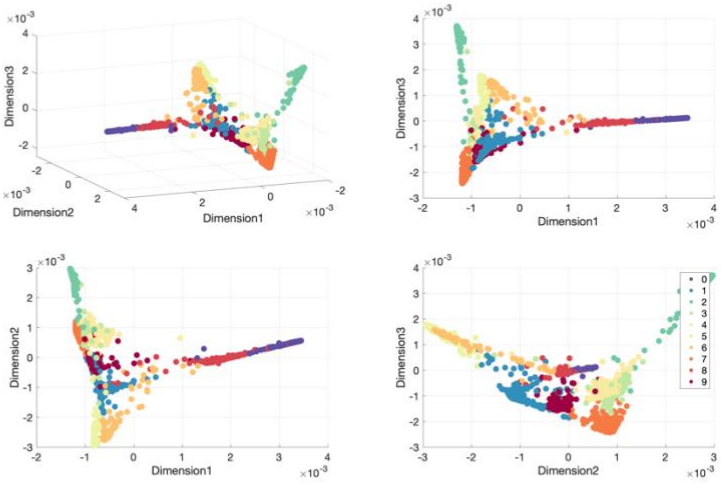


Figure 16 Batch MKL Output space for the Digits dataset

The space obtained with the batch approach displays the characteristic shape of Laplacian Eigenmaps, positioning the different digit classes along elongated branches. The separability appears to be good and observations within the same class are clustered together.

I. Multi-class growth Experiments

We present the result of our experiment where the seed space is initially trained with varying quantities of observations—50, 150, and 190—from each of the 200-digit observations of each digit class. Following this, we then expanded the seed space using the residual observations—150, 50, and 10, respectively. The visual representation of these results can be seen in Figures 17 through 20.

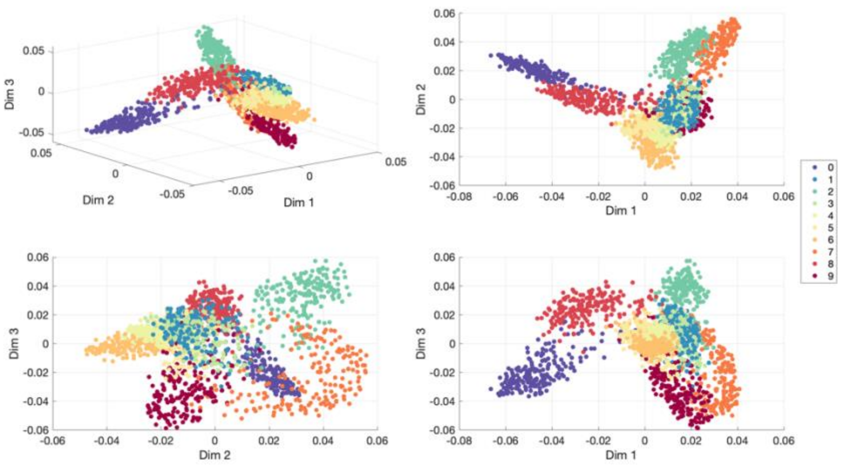


Figure 17 Output Space trained on the first observation of each digit and grown with the remaining 199 observations. The neighborhood size used was 5.

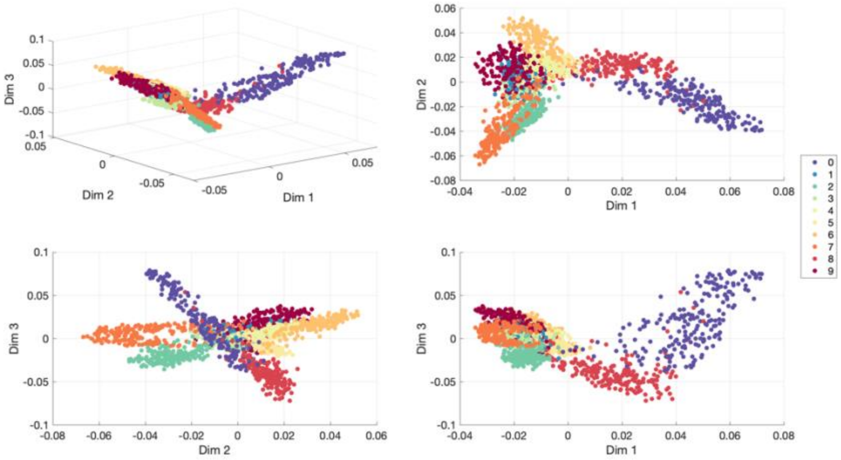


Figure 18 Output Space trained on the first 10 observations of each digit and grown with the remaining 190 observations. The neighborhood size used was 5.

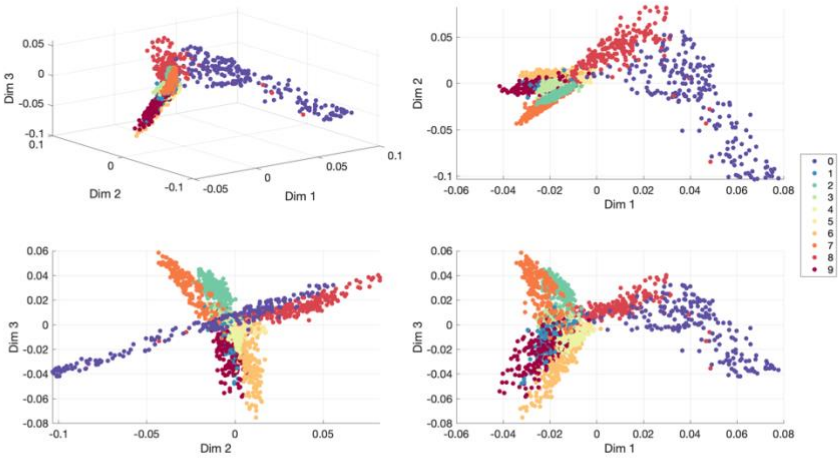


Figure 19 Output Space trained on the first 100 observations of each digit and grown with the remaining 100 observations. The neighborhood size used was 3.

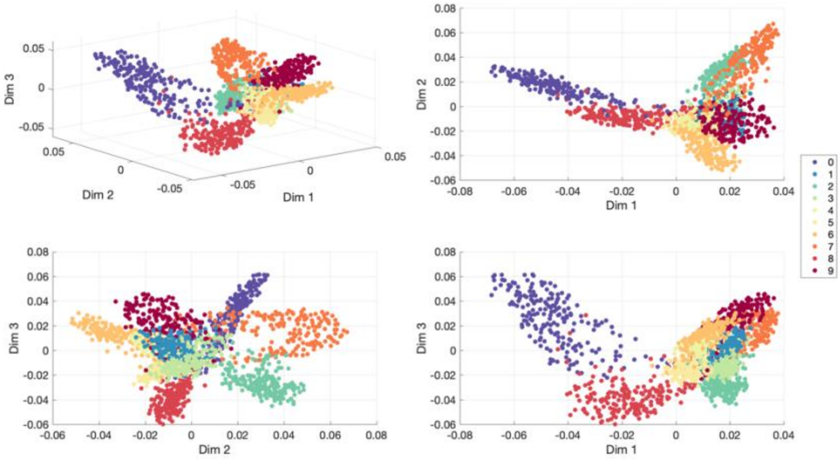


Figure 20 Output Space trained on the first 150 observations of each digit and grown with the remaining 50 observations. The neighborhood size used was 3.

Our incremental method consistently demonstrates good performance across all experiments. Remarkably, it properly positions new samples, resulting in a space that clusters together observations from the same class and separates different ones.

This approach maintains its effectiveness even under conditions of significantly constrained initial data sets, where the seed spaces were exceedingly small, comprised of merely 10 observations - with one observation representing one of the 10 digits.

II. Class Addition Experiments

Our study presents the outcomes of various training strategies on the seed space, specifically using 5-, 7-, and 9-digit classes, and subsequently expanding the space with the remaining 5, 3, and 1 digit classes respectively. In Figures 21, 23, and 25, we display the results obtained from employing a batch MKL-DR training approach on all 200 observations of the initial five digits (ranging from 0 to 4). Following this, we employed our suggested incremental MKL-DR (iMKL-DR) approach to expand the seed space with the remaining digits (5 to 9). The displayed results thus reflect the comparative effectiveness of our training and expansion methods in Figures 22, 24, 26.

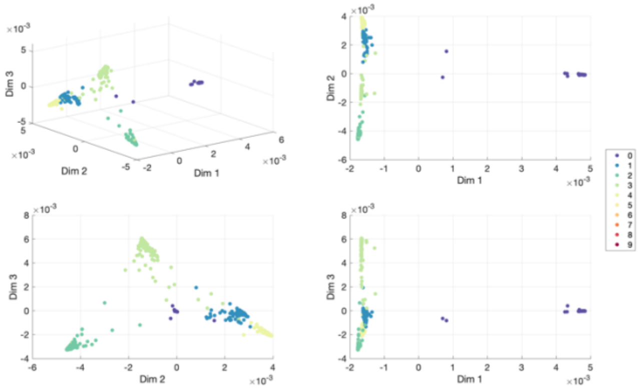


Figure 21 Output Space of MKL-DR trained on the 200 observations of the first 5 digits

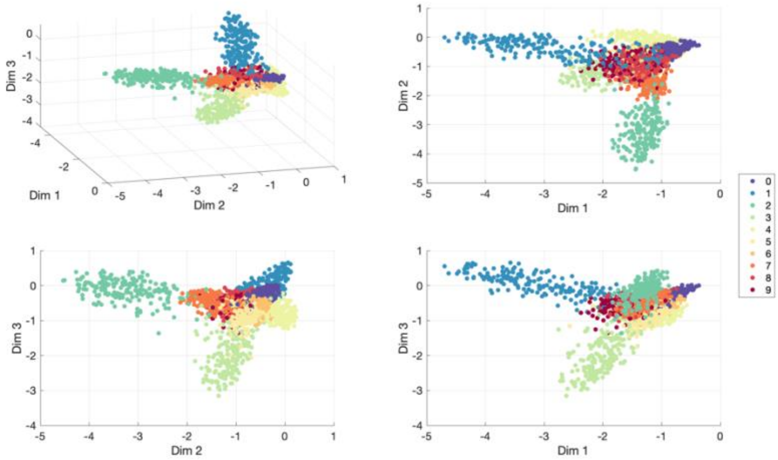


Figure 22 Output Space trained on the first 5 digits and grown with the remaining 5 digits

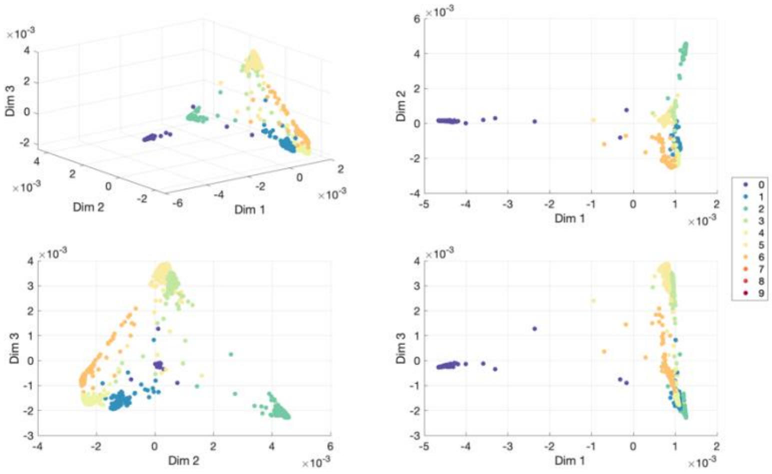


Figure 23 Output Space of MKL-DR trained on the 200 observations of the first 7 digits

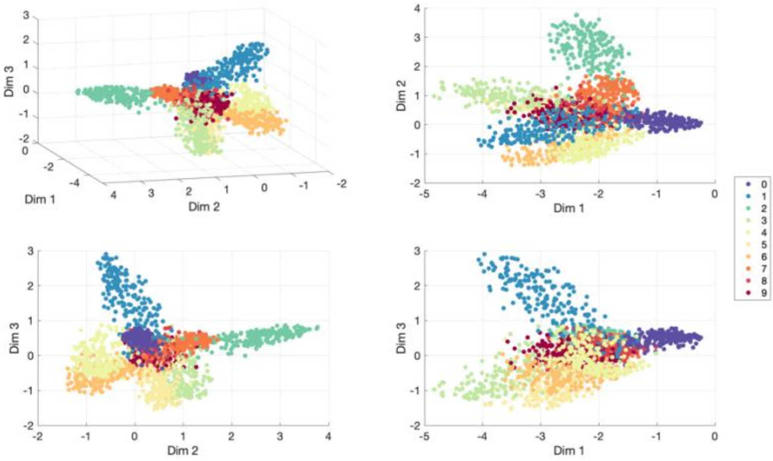


Figure 24 Output Space trained on the first 7 digits and grown with the remaining 3 digits

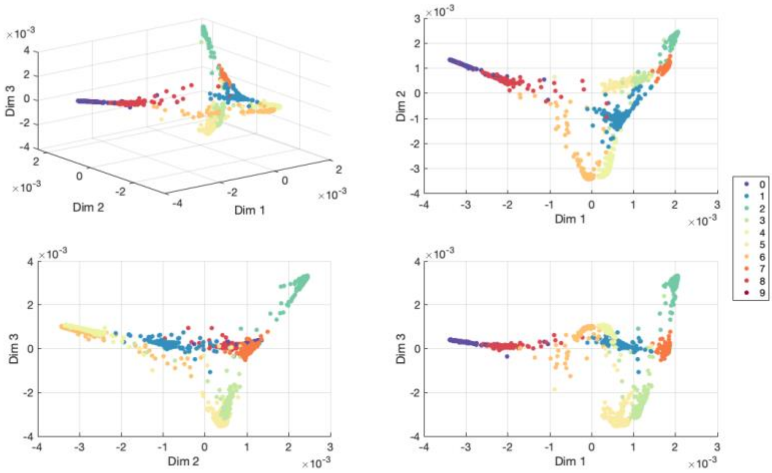


Figure 25 Output Space of MKL-DR trained on the 200 observations of the first 9 digits

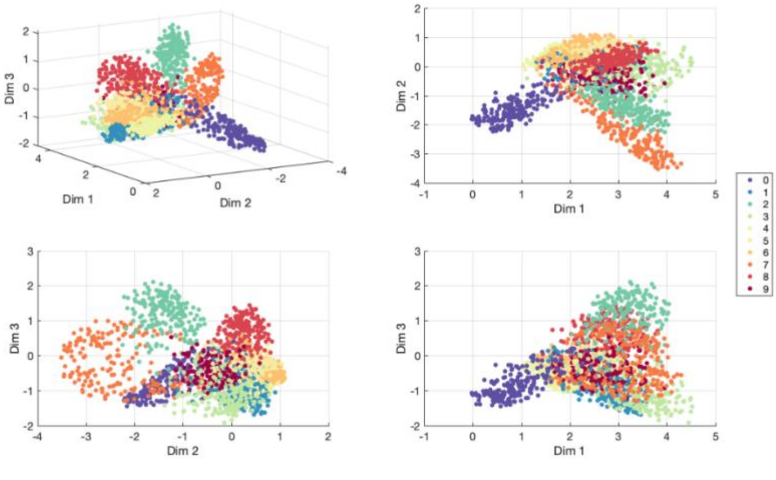


Figure 26 Output Space trained on the first 9 digits and grown with the remaining digit

Consistent with the batch MKL space on the full dataset, our findings indicate that when training is conducted using smaller subsets of data, containing only a few classes, the observations from different digit classes are arranged in the same branching pattern.

In contrast, our incremental approach yields distributions that are more dispersed, appearing more like lobes rather than branches. Nonetheless, our method still manages to preserve both the separability between classes and the cohesion within each class.

Therefore, in the context of this dataset, we can assert that our incremental methodology is properly integrating unseen classes into the existing embedding. Thus, our algorithm effectively accommodates new information, successfully maintaining the crucial attributes of class distinction and intra-class unity.

III. Sensitivity analysis for the seed and neighborhood size

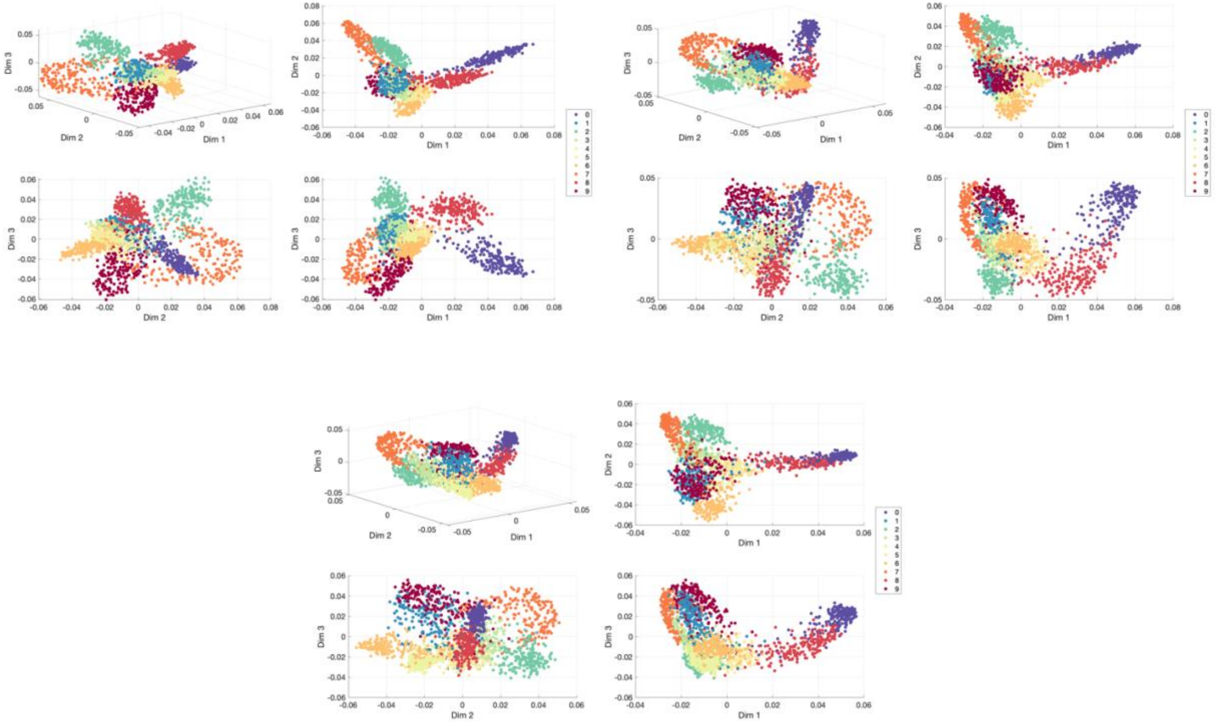


Figure 27 Effect of Neighborhood size using a seed space trained on 1 observation of each digit.

Top Left: KNN value of 1; Top Right: KNN value of 50; Bottom: KNN of 150

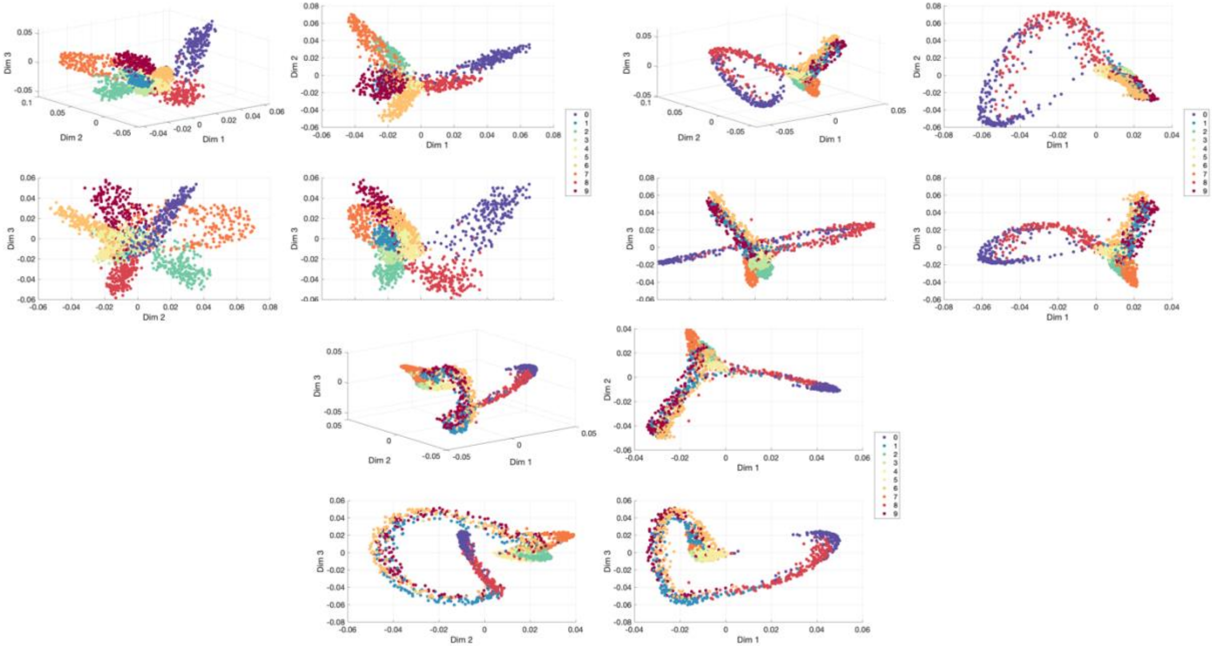


Figure 28 Effect of Neighborhood size using a seed space trained on 50 observations of each digit.

Top Left: KNN value of 1; Top Right: KNN value of 50; Bottom: KNN of 150

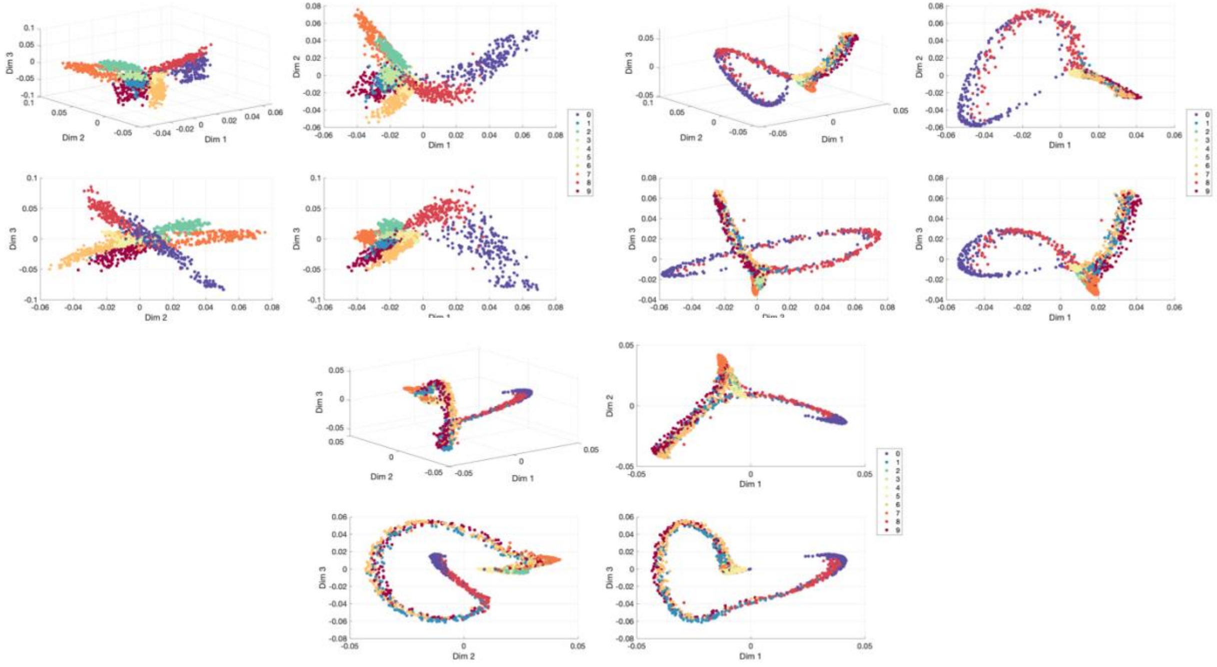


Figure 29 Effect of Neighborhood size using a seed space trained on 100 observations of each digit.

Top Left: KNN value of 1; Top Right: KNN value of 50; Bottom: KNN of 150

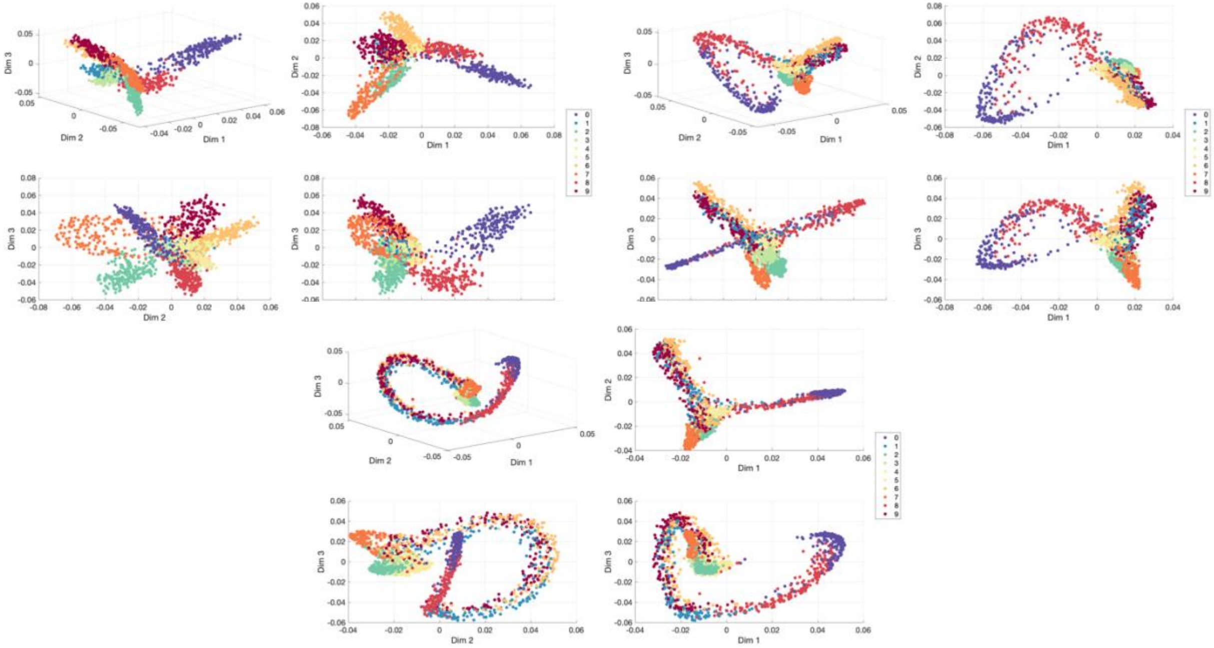


Figure 30 Effect of Neighborhood size using a seed space trained on 150 observations of each digit.

Top Left: KNN value of 1; Top Right: KNN value of 50; Bottom: KNN of 150

This last study reveals important relationships between the size of the initial group of observations, known as the seed space, and the neighborhood size when using an incremental extension to unsupervised multiple kernel learning (MKL). We visualize the results in Figures 27, 28, 29, 30.

When the seed space is small (10 observations), there is a lot of flexibility in choosing the neighborhood size. Both very large and small neighborhoods can be used without impacting the overall structure of the space, suggesting the method is stable when the seed space is small. However, as the seed space grows, careful selection of the neighborhood size becomes necessary. Larger seed spaces require the use of smaller neighborhoods. The trend continues to the point where, if the seed space is significantly large, very small neighborhoods must be used. If this adjustment is not made, the space collapses and observations from different classes begin to mix, which is not desirable. This mixing likely occurs because batch MKL tends to create very separate clusters, and the incremental MKL has trouble adjusting the positions of a large number of distinct observations when new ones are added.

The results highlight the balance needed when using incremental MKL. It is crucial to carefully adjust the parameters of seed space size and neighborhood size to get optimal embeddings, as these factors heavily influence each other. More research may help us understand these dynamics better and potentially develop a way to automatically or adaptively set these parameters.

- Hypertensive dataset

In the same way as before, we first provide the results of running batch MKL on the full dataset to allow for comparisons. The visualization corresponds to Figure 31.

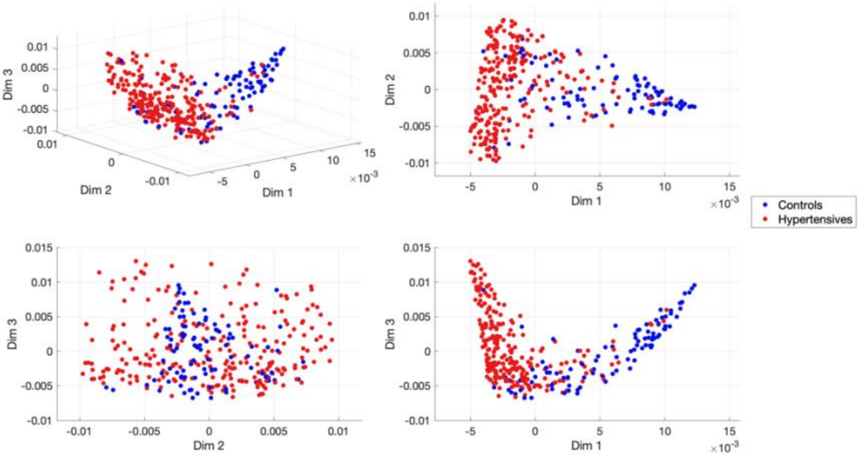


Figure 31 Batch MKL Output space for the Hypertensive dataset

We observe a clear separation between classes and a defined spectrum between disease and normalcy. The hypertense patients display a broader spectrum while the controls are concentrated in a corner.

I. Multi-class growth Experiments

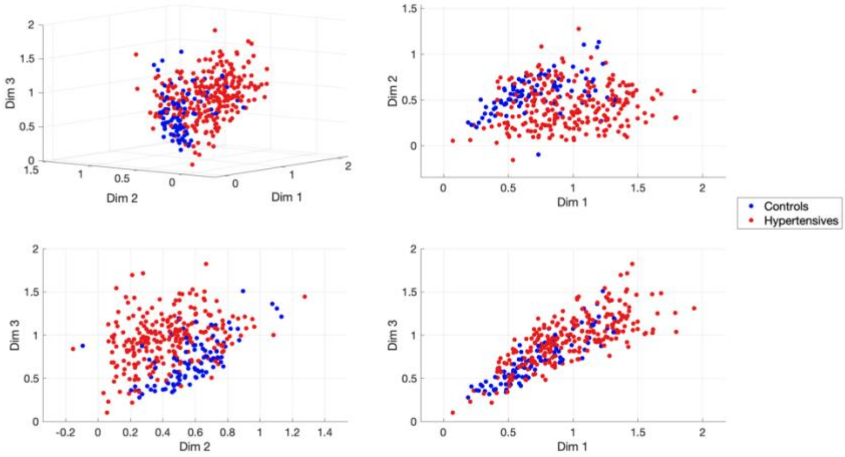


Figure 32 Output Space trained on the first 5 observations of each class and grown with the remaining ones

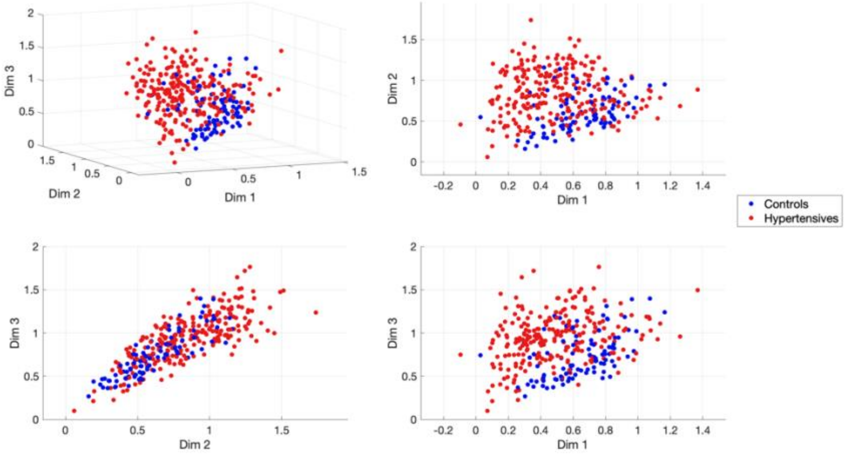


Figure 33 Output Space trained on the first 30 observations of each class and grown with the remaining ones

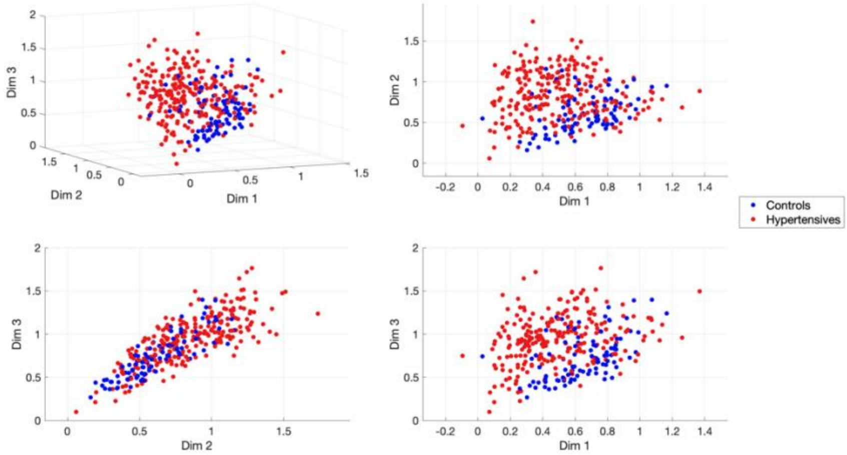


Figure 34 Output Space trained on the first 85 observations of each class and grown with the remaining ones

We observe a subtle separation between classes regardless of the size of the seed space in Figures 32, 33, 34. Training with 5 observations from each class resulted in the worst separability, while training with 85 observations of each class had the best discriminative ability. There is still a spectrum from normalcy to disease, but it is much less defined than in the space obtained with batch MKL.

II. Class Addition Experiments

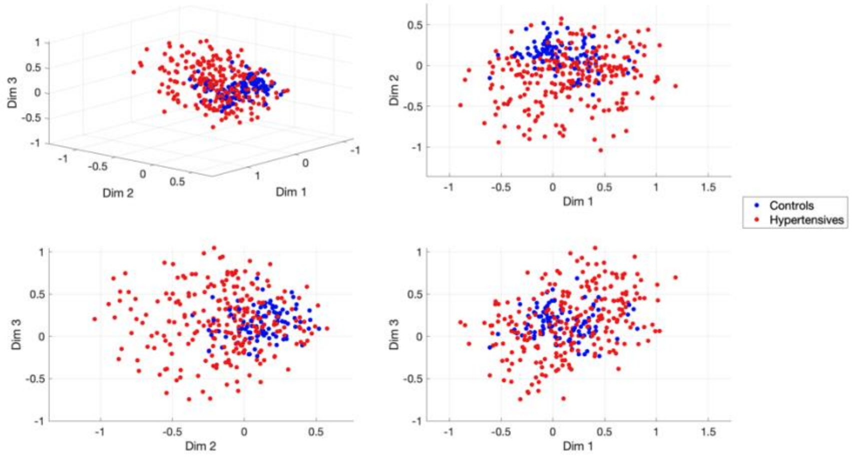


Figure 35 Output space obtained training the seed with hypertensives and growing with controls

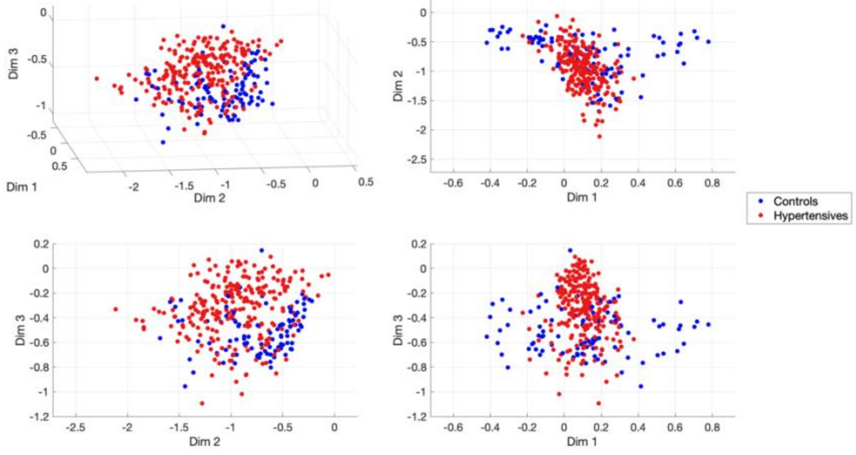


Figure 36 Output space obtained training the seed with controls and growing with hypertense patients

Our results vary noticeably when applying different strategies in the seed space training process, as shown in Figures 35 and 36.

Initially, we trained the seed space using one class, then expanded it with the other.

Specifically, when we started with hypertensive subjects for training and subsequently introduced control subjects, the distinctions between the two groups were not evident. In this scenario, the controls ended up closely intermingled within the hypertensive population, thereby failing to delineate clear divisions between the two groups.

In contrast, using control subjects for initial training and then expanding the seed space with hypertensive subjects yielded a somewhat different outcome. The generated embedding in this case showed some level of separation between the two classes. However, it is important to note that this separation was not as good as the one obtained from the batch approach. In the context of this dataset, the iMKL-DR results were inferior compared to the batch results.

4.5 Discussion

To our knowledge, this is the first work to provide a purely incremental extension of unsupervised MKL-DR. The results obtained show promise for accommodating larger datasets to the unsupervised MKL scenario. Mainly due to their capacity to iteratively add an increasing number of observations and position them correctly by similarity.

Nonetheless, we observed that it is crucial to assess the a priori similarity between the existing seed space and the new samples that will be used to grow it. This was especially crucial in the clinical dataset where training on one class and adding the complementary one yielded poor performance, most probably due to missing "anchor" points to which newly coming samples could latch on to. In essence, if the newly coming samples differ greatly to the ones in the seed space, there is a risk for the incremental approach to fall in a bad solution and collapse all points in an undesired region of the embedding.

We acknowledge that our work represents a somewhat preliminary exploration into incrementally extending unsupervised multiple kernel learning for dimensionality reduction. As such, there are several limitations that provide opportunities for future work. First, more extensive validation is needed to fully evaluate the proposed approach. Specifically, it should be tested on a diverse array of datasets and compared against other state-of-the-art dimensionality reduction techniques on relevant metrics. This will provide a more comprehensive understanding of when and how the proposed approach is most beneficial.

Second, applying a broad battery of quantitative measures could reveal further insights into the trade-offs made during dimensionality reduction. Key metrics to consider include those that quantify information loss, preservation of local data structure, and computational cost analyses.

Lastly, if possible, fMKL could be implemented to reach even faster training speeds when training the seed space. This would provide a faster formulation of the algorithm.

5. Conclusions

This thesis has successfully employed information fusion techniques to address critical problems in healthcare and clinical research. The four projects completed demonstrate the versatility of classical statistics, supervised learning, and unsupervised learning in extracting insights from complex medical data.

Regarding the improvement of efficiency in the clinical trial setting, unsupervised MKL effectively learned latent representations from heterogeneous population data. This technique shows promise for creating synthetic control arms in clinical trials, pending improvements in computational scalability. The ML-derived selection of controls currently provides reasonably accurate matches, but continued continuous updating would be needed to account for population shifts over time.

When dealing with the prediction of SGA pregnancies, we successfully leveraged gradient boosting models to predict SGA across the two distinct cohorts of Barcelona and Karachi. The comparative analysis revealed differing predictive capacities of features like maternal factors, biometry, and Doppler measurements between populations. This indicates a need to adjust models to the specific cohort when transferring algorithms across geographies. Larger sample sizes are also required to improve detection of rare outcomes like preterm birth.

Social and data science methods were able to provide insights into the differences between clubfoot and control families in India.

While illuminating, the study was limited by its cross-sectional nature as opposed to a study with longitudinal data collection, which hindered the possibility of strong causal explanations. We were also aware of potential response biases in the clubfoot affected families. Nonetheless, we were able to create and deploy refined, culturally adapted questionnaires, which we are confident will provide rich insights into patient outcomes when we use them prospectively. All in all, the work validates the holistic assessment of patients to understand intervention effects.

Finally, our technical contribution proposed modifications to increase the scalability of unsupervised multiple kernel learning. This incremental learning approach shows initial promise in accommodating larger datasets. Further validation on diverse datasets is needed to fully demonstrate its advantages over existing techniques.

In conclusion, this thesis has shown the potential of data science methodologies to extract insights from complex medical data. While limitations exist, the projects demonstrate a promising line of research for these techniques in critical healthcare applications pending refinements in cohort size, model validation, and questionnaire designs.

Bibliography

1. World Health Organization. Health workforce [Internet]. Geneva: World Health Organization; [cited 2023 Jul 13]. Available from: https://www.who.int/health-topics/health-workforce#tab=tab_1.

2. Zhang A, Xing L, Zou J, Wu JC. Shifting machine learning for healthcare from development to deployment and from models to data. Nat Biomed Eng. 2022 Jul 4;6(12):1330–45.

3. Albahri AS, Duhaim AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. Information Fusion. 2023 Aug;96:156–91.

4. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, et al. Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association. Circulation [Internet]. 2019 Mar 5 [cited 2023 Jul 17];139(10). Available from: https://www.ahajournals.org/doi/10.1161/CIR.0000000000000659

5. Fleming TP, Watkins AJ, Velazquez MA, Mathers JC, Prentice AM, Stephenson J, et al. Origins of lifetime health around the time of conception: causes and consequences. Lancet. 2018 May 5;391(10132):1842–52.

6. Martin L, Hutchens M, Hawkins C, Radnov A. How much do clinical trials cost? Nat Rev Drug Discov. 2017 Jun;16(6):381–2.

7. Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, Robinson JG, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory: Table 1. J Am Med Inform Assoc. 2013 Dec;20(e2):e226–31.

8. WHO. Every newborn: an action plan to end preventable deaths. 2014.

9. Catalyst NE. What is value-based healthcare?. NEJM Catalyst. 2017 Jan 1;3(1).

10. Lauer MS, Gordon D, Wei G, Pearson G. Efficient design of clinical trials and epidemiological research: is it possible? Nature Reviews Cardiology. 2017 Aug 1;14(8):493–502.

11. Cameron D, Willoughby C, Messer D, Lux M, Aitken M, Getz K. Assessing Participation Burden in Clinical Trials: Introducing the Patient Friction Coefficient. Clinical Therapeutics. 2020 Aug 1;42(8):e150–9.

12. Borno H, Siegel A, Ryan C. The problem of representativeness of clinical trial participants: understanding the role of hidden costs. J Health Serv Res Policy. 2016 Jul;21(3):145–6.

13. Harrer S, Shah P, Antony B, Hu J. Artificial Intelligence for Clinical Trial Design. Trends in Pharmacological Sciences. 2019 Aug 1;40(8):577–91.

14. U.S. Food and Drug Administration [Internet]. 2018 [cited 2023 Jun 16]. How Simulation Can Transform Regulatory Pathways. Available from: https://www.fda.gov/science-research/about-science-research-fda/how-simulation-can-transform-regulatory-pathways

15. Ghadessi M, Tang R, Zhou J, Liu R, Wang C, Toyoizumi K, et al. A roadmap to using historical controls in clinical trials – by Drug Information Association Adaptive Design Scientific Working Group (DIA-ADSWG). Orphanet Journal of Rare Diseases. 2020 Mar 12;15(1):69.

16. Sanchez-Martinez S, Duchateau N, Erdei T, Fraser AG, Bijnens BH, Piella G. Characterization of myocardial motion patterns by unsupervised multiple kernel learning. Medical Image Analysis. 2017;35:70–82.

17. Cikes M, Sanchez-Martinez S, Claggett B, Duchateau N, Piella G, Butakoff C, et al. Machine learning-based phenogrouping in

heart failure to identify responders to cardiac resynchronization therapy. European Journal of Heart Failure. 2019;21(1):74–85.

18. Shah AM, Cheng S, Skali H, Wu J, Mangion JR, Kitzman D, et al. Rationale and design of a multicenter echocardiographic study to assess the relationship between cardiac structure and function and heart failure risk in a biracial cohort of community-dwelling elderly persons: the Atherosclerosis Risk in Communities study. Circ Cardiovasc Imaging. 2014 Jan;7(1):173–81.

19. THE ARIC INVESTIGATORS. THE ATHEROSCLEROSIS RISK IN COMMUNIT (ARIC) STUDY: DESIGN AND OBJECTIVES. American Journal of Epidemiology. 1989 Apr 1;129(4):687–702.

20. Moss AJ, Brown MW, Cannom DS, Daubert JP, Estes M, Foster E, et al. Multicenter automatic defibrillator implantation trial-cardiac resynchronization therapy (MADIT-CRT): design and clinical protocol. Ann Noninvasive Electrocardiol. 2005 Oct;10(4 Suppl):34–43.

21. Moss AJ, Hall WJ, Cannom DS, Klein H, Brown MW, Daubert JP, et al. Cardiac-resynchronization therapy for the prevention of heart-failure events. N Engl J Med. 2009 Oct 1;361(14):1329–38.

22. Shah AM, Shah SJ, Anand IS, Sweitzer NK, O’Meara E, Heitner JF, et al. Cardiac structure and function in heart failure with preserved ejection fraction: baseline findings from the echocardiographic study of the Treatment of Preserved Cardiac Function Heart Failure with an Aldosterone Antagonist trial. Circ Heart Fail. 2014 Jan;7(1):104–15.

23. Pitt B, Pfeffer MA, Assmann SF, Boineau R, Anand IS, Claggett B, et al. Spironolactone for Heart Failure with Preserved Ejection Fraction. New England Journal of Medicine. 2014 Apr 10;370(15):1383–92.

24. Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, et al. Recommendations for cardiac chamber

quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. J Am Soc Echocardiogr. 2015 Jan;28(1):1-39.e14.

25. Sanchez-Martinez S, Duchateau N, Erdei T, Kunszt G, Aakhus S, Degiovanni A, et al. Machine Learning Analysis of Left Ventricular Function to Characterize Heart Failure With Preserved Ejection Fraction. Circulation: Cardiovascular Imaging [Internet]. 2018 Apr [cited 2019 Sep 19];11(4). Available from: https://www.ahajournals.org/doi/10.1161/CIRCIMAGING.117.007138

26. Loncaric F, Castellote PMM, Sanchez-Martinez S, Fabijanovic D, Nunno L, Mimbrero M, et al. Automated Pattern Recognition in Whole-Cardiac Cycle Echocardiographic Data: Capturing Functional Phenotypes with Machine Learning. Journal of the American Society of Echocardiography. 2021 Nov 1;34(11):1170–83.

27. Lin YY, Liu TL, Fuh CS. Multiple kernel learning for dimensionality reduction. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2011;33(6):1147–60.

28. Namasivayam M. Machine Learning in Cardiac Imaging: Exploring the Art of Cluster Analysis. J Am Soc Echocardiogr. 2021 Aug;34(8):913–5.

29. Pfeffer MA, Claggett B, Assmann SF, Boineau R, Anand IS, Clausell N, et al. Regional variation in patients and outcomes in the Treatment of Preserved Cardiac Function Heart Failure With an Aldosterone Antagonist (TOPCAT) trial. Circulation. 2015 Jan 6;131(1):34–42.

30. Williams C, Seeger M. Using the Nyström Method to Speed Up Kernel Machines. In: Leen T, Dietterich T, Tresp V, editors. Advances in Neural Information Processing Systems [Internet]. MIT Press; 2000. Available from: https://proceedings.neurips.cc/paper_files/paper/2000/file/19de10adbaa1b2ee13f77f679fa1483a-Paper.pdf

31. Malik ZK, Hussain A, Wu J. An online generalized eigenvalue version of Laplacian Eigenmaps for visual big data. Neurocomputing. 2016;173:127–36.

32. Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. European heart journal. 2012 Jan;33(2):176–82.

33. Liu L, Oza S, Hogan D, Chu Y, Perin J, Zhu J, et al. Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals. The Lancet. 2016 Dec;388(10063):3027–35.

34. Battaglia FC, Lubchenco LO. A practical classification of newborn infants by weight and gestational age. The Journal of Pediatrics. 1967 Aug;71(2):159–63.

35. Lundgren EM, Cnattingius S, Jonsson B, Tuvemo T. Intellectual and Psychological Performance in Males Born Small for Gestational Age With and Without Catch-Up Growth. Pediatr Res. 2001 Jul;50(1):91–6.

36. Finken MJJ, Van Der Steen M, Smeets CCJ, Walenkamp MJE, De Bruin C, Hokken-Koelega ACS, et al. Children Born Small for Gestational Age: Differential Diagnosis, Molecular Genetic Evaluation, and Implications. Endocrine Reviews. 2018 Dec 1;39(6):851–94.

37. Bhutta ZA, Das JK, Bahl R, Lawn JE, Salam RA, Paul VK, et al. Can available interventions end preventable deaths in mothers, newborn babies, and stillbirths, and at what cost? The Lancet. 2014 Jul;384(9940):347–70.

38. Wright S, Mathieson K, Brearley L, Jacobs S, Holly L, Wickremasinghe R, et al. Ending newborn deaths: ensuring every baby survives. 2014.

39. Abdullah MA, Mukhtar F, Wazir S, Gilani I, Gorar Z, Shaikh BT. The health workforce crisis in Pakistan: a critical review and the way forward. World Health Popul. 2014;15(3):4–12.

40. King VJ, Bennet L, Stone PR, Clark A, Gunn AJ, Dhillon SK. Fetal growth restriction and stillbirth: Biomarkers for identifying at risk fetuses. Front Physiol. 2022 Aug 19;13:959750.

41. Crispi F, Sepúlveda-Martínez Á, Crovetto F, Gómez O, Bijnens B, Gratacós E. Main Patterns of Fetal Cardiac Remodeling. Fetal Diagn Ther. 2020;47(5):337–44.

42. Garcia-Canadilla P, Crispi F, Cruz-Lemini M, Triunfo S, Nadal A, Valenzuela-Alcaraz B, et al. Patient-specific estimates of vascular and placental properties in growth-restricted fetuses based on a model of the fetal circulation. Placenta. 2015 Sep;36(9):981–9.

43. Garcia-Canadilla P, Rudenick PA, Crispi F, Cruz-Lemini M, Palau G, Camara O, et al. A Computational Model of the Fetal Circulation to Quantify Blood Redistribution in Intrauterine Growth Restriction. Gefen A, editor. PLoS Comput Biol. 2014 Jun 12;10(6):e1003667.

44. Dunn L, Sherrell H, Kumar S. Review: Systematic review of the utility of the fetal cerebroplacental ratio measured at term for the prediction of adverse perinatal outcome. Placenta. 2017 Jun;54:68–75.

45. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? BMJ Glob Health. 2018 Aug;3(4):e000798.

46. Garcia-Canadilla P, Sanchez-Martinez S, Crispi F, Bijnens B. Machine Learning in Fetal Cardiology: What to Expect. Fetal Diagnosis and Therapy. 2020;47(5):363–72.

47. The investigation and management of the small for gestational age fetus. Royal College of Obstetricians and Gynaecologists. Updated January 2014. Accessed November 11, 2021. https://www.rcog.org.uk/globalassets/documents/guidelines/gtg_31.pdf.

48. Crovetto F, Crispi F, Casas R, Martín-Asuero A, Borràs R, Vieta E, et al. Effects of Mediterranean Diet or Mindfulness-Based Stress Reduction on Prevention of Small-for-Gestational Age Birth Weights in Newborns Born to At-Risk Pregnant Individuals: The IMPACT BCN Randomized Clinical Trial. JAMA. 2021 Dec 7;326(21):2150.

49. Hoodbhoy Z, Hasan B, Jehan F, Bijnens B, Chowdhury D. Machine learning from fetal flow waveforms to predict adverse perinatal outcomes: A study protocol. Gates Open Research. 2018;2(May):1–13.

50. Papageorghiou AT, Kennedy SH, Salomon LJ, Altman DG, Ohuma EO, Stones W, et al. The INTERGROWTH-21st fetal growth standards: toward the global integration of pregnancy and pediatric care. Am J Obstet Gynecol. 2018 Feb;218(2S):S630–40.

51. Figueras F, Meler E, Iraola A, Eixarch E, Coll O, Figueras J, et al. Customized birthweight standards for a Spanish population. European Journal of Obstetrics & Gynecology and Reproductive Biology. 2008 Jan;136(1):20–4.

52. Committee Opinion No 579: Definition of Term Pregnancy. Obstetrics & Gynecology. 2013 Nov;122(5):1139–40.

53. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. San Francisco California USA: ACM; 2016 [cited 2023 Jun 26]. p. 785–94. Available from: https://dl.acm.org/doi/10.1145/2939672.2939785

54. Blencowe H, Krasevec J, De Onis M, Black RE, An X, Stevens GA, et al. National, regional, and worldwide estimates of low birthweight in 2015, with trends from 2000: a systematic analysis. The Lancet Global Health. 2019 Jul;7(7):e849–60.

55. Punyapet P, Suwanrath C, Chainarong N, Sawaddisan R, Vichitkunakorn P. Predictors of adverse perinatal outcomes in fetal growth restriction using a combination of maternal clinical

factors and simple ultrasound parameters. Intl J Gynecology & Obste. 2023 Mar 3;ijgo.14721.

56. Kuhle S, Maguire B, Zhang H, Hamilton D, Allen AC, Joseph KS, et al. Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: a retrospective cohort study. BMC Pregnancy Childbirth. 2018 Dec;18(1):333.

57. Iwama N, Obara T, Ishikuro M, Murakami K, Ueno F, Noda A, et al. Risk scores for predicting small for gestational age infants in Japan: The TMM birthree cohort study. Sci Rep. 2022 May 26;12(1):8921.

58. Lian C, Wang Y, Bao X, Yang L, Liu G, Hao D, et al. Dynamic prediction model of fetal growth restriction based on support vector machine and logistic regression algorithm. Front Surg. 2022 Sep 23;9:951908.

59. Anderson NH, Sadler LC, McKinlay CJD, McCowan LME. INTERGROWTH-21st vs customized birthweight standards for identification of perinatal mortality and morbidity. American Journal of Obstetrics and Gynecology. 2016 Apr;214(4):509.e1-509.e7.

60. Poon LCY, Tan MY, Yerlikaya G, Syngelaki A, Nicolaides KH. Birth weight in live births and stillbirths: Birth weight in live births and stillbirths. Ultrasound Obstet Gynecol. 2016 Nov;48(5):602–6.

61. Owen RM, Capper B, Lavy C. Clubfoot treatment in 2015: a global perspective. BMJ Global Health. 2018 Sep 1;3(4).

62. Hussain H, Burfat AM, Samad L, Jawed F, Chinoy MA, Khan MA. Cost-Effectiveness of the Ponseti Method for Treatment of Clubfoot in Pakistan. World Journal of Surgery. 2014 Apr 8;38(9):2217–22.

63. Harmer L, Rhatigan J. Clubfoot Care in Low-Income and Middle-Income Countries: From Clinical Innovation to a Public Health Program. World Journal of Surgery. 2014 Apr 1;38(4):839–48.

64. Francesc Malagelada, Malagelada F, Mayet S, Firth G, Ramachandran M, Manoj Ramachandran. The impact of the Ponseti treatment method on parents and caregivers of children with clubfoot: a comparison of two urban populations in Europe and Africa. Journal of Children’s Orthopaedics. 2016 Apr 1;10(2):101–7.

65. Iqbal MS, Dubey R, Thakur K, Katiyar S, Prasad M. Assessment of awareness and barriers to clubfoot treatment in the Indian scenario. J Family Med Prim Care. 2021 Nov;10(11):4229–35.

66. Sharaf Sheik-Ali, Navarro SM, Keil EJ, Evan Keil, Walter Johnson, Chris Lavy. The Health Determinants of Accessibility to Clubfoot Treatment in LMICs: A Global Exploration of Barriers and Solutions. International Journal of MCH and AIDS. 2021 Dec 2;10(2):241–50.

67. Drew S, Lavy C, Gooberman-Hill R. What factors affect patient access and engagement with clubfoot treatment in low- and middle-income countries? Meta-synthesis of existing qualitative studies using a social ecological model. Trop Med Int Health. 2016 May;21(5):570–89.

68. Evans A, Chowdhury M, Karimi L, Rouf A, Uddin S, Haque O. Factors Affecting Parents to ‘Drop-Out’ from Ponseti Method and Children’s Clubfoot Relapse. Orthopedic Research Online Journal. 2020 Jan 7;6(1):1–9.

69. Clubfoot India. Available from: https://clubfootindia.in/. [Accessed: 2023 July 17].

70. International Consortium for Health Outcomes Measurement. Congenital Heart Disease [Internet]. Boston: ICHOM; 2023 [cited 2023 Jul 20]. Available from: https://connect.ichom.org/patient-centered-outcome-measures/congenital-heart-disease/.

71. International Consortium for Health Outcomes Measurement. Congenital Upper Limb Anomalies [Internet]. Boston: ICHOM; 2023 [cited 2023 Jul 20]. Available from:

https://connect.ichom.org/patient-centered-outcome-measures/congenital-upper-limb-anomalies/.

72. Saxena S, Carlson D, Rex Billington, Billington R, Orley J. The WHO quality of life assessment instrument (WHOQOL-Bref): The importance of its items for cross-cultural research. Quality of Life Research. 2001 Jan 1;10(8):711–21.

73. Suranjan Majumder. Socioeconomic status scales: Revised Kuppuswamy, BG Prasad, and Udai Pareekh's scale updated for 2021. Journal of family medicine and primary care. 2021 Jan 1;

74. Thornicroft G, Brohan E, Rose D, Sartorius N, Sartorius N, Leese M. Global pattern of experienced and anticipated discrimination against people with schizophrenia: a cross-sectional survey. The Lancet. 2009 Jan 31;373(9661):408–15.

75. Jain A, Vishwanathan SVN, Varma M. SPF-GMKL: generalized multiple kernel learning with a million kernels. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining [Internet]. Beijing China: ACM; 2012 [cited 2023 Jul 17]. p. 750–8. Available from: https://dl.acm.org/doi/10.1145/2339530.2339648

76. Rakotomamonjy A, Bach FR, Canu S, Grandvalet Y. SimpleMKL. Journal of Machine Learning Research. 2008;9(83):2491–521.

77. Meanti G, Carratino L, Rosasco L, Rudi A. Kernel Methods Through the Roof: Handling Billions of Points Efficiently. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2020. p. 14410–22. Available from: https://proceedings.neurips.cc/paper_files/paper/2020/file/a59afb1b7d82ec353921a55c579ee26d-Paper.pdf

78. Xiang L, Zhao G, Li Q, Hao W, Li F. TUMK-ELM: A Fast Unsupervised Heterogeneous Data Learning Approach. IEEE Access. 2018;6:35305–15.

79. Giang TT, Nguyen TP, Nguyen TQV, Tran DH. fMKL-DR: A Fast Multiple Kernel Learning Framework with Dimensionality Reduction. In: Huynh VN, Inuiguchi M, Tran DH, Denoeux T, editors. Integrated Uncertainty in Knowledge Modelling and Decision Making [Internet]. Cham: Springer International Publishing; 2018 [cited 2023 Jul 17]. p. 153–65. (Lecture Notes in Computer Science; vol. 10758). Available from: http://link.springer.com/10.1007/978-3-319-75429-1_13

80. Li H, Jiang H, Barrio R, Liao X, Cheng L, Su F. Incremental manifold learning by spectral embedding methods. Pattern Recognition Letters. 2011;32(10):1447–55.

81. Duin R. Multiple Features [Internet]. UCI Machine Learning Repository; 1998 [cited 2023 Jul 17]. Available from: https://archive.ics.uci.edu/dataset/72