Machine learning to support exploring and exploiting real-world clinical longitudinal data

Mariana Nogueira

TESI DOCTORAL UPF / 2020

THESIS SUPERVISORS
Bart Bijnens, Gemma Piella, Mathieu De Craene

Department of Engineering and Information and Communication Technologies

Universitat Pompeu Fabra Barcelona

Acknowledgements

This thesis marks the conclusion of a long journey, and I have many people to thank for helping me reach this point.

Firstly, I would like to thank my supervisors, Bart, Gemma and Mathieu. All this would simply not have been possible without their guidance and support. I feel very lucky to have had the opportunity to closely collaborate with and learn from them.

I would also like to thank Devyani and Olufemi for participating in this work with their invaluable contributions.

A very special thanks to my friends and colleagues from Philips and UPF, for always being up for an exchange of ideas, and for the much-needed decompression moments over beers. A special mention to my *CFX* mates.

Amidst the adventures of flatsharing, I was lucky enough to make great friends, who made good times better and harder times easier to cope with. I would especially like to thank my *CLJT Suresnes*, *Padilla* and *Dos de Maig* mates.

A very special thanks to my friends from Portugal, whom I could always count on despite the distance.

Lastly, I would like to thank my family, my most important support system, for always and unconditionally being there, through good times and harder times.

Abstract

Following-up on patient evolution by reacquiring the same measurements over time (longitudinal data) is a crucial component in clinical care dynamics, as it creates opportunity for timely decision making in preventing adverse outcome. It is thus important that clinicians have proper longitudinal analysis tools at their service. Nonetheless, most traditional longitudinal analysis tools have limited applicability if data are (1) not highly standardized or (2) very heterogeneous (e.g. images, signal, continuous and categorical variables) and/or high-dimensional. These limitations are extremely relevant, as both scenarios are prevalent in routine clinical practice.

The aim of this thesis is the development of tools that facilitate the integration and interpretation of complex and nonstandardized longitudinal clinical data. Specifically, we explore approaches based on unsupervised dimensionality reduction, which allow the integration of complex longitudinal data and their representation as low-dimensional yet clinically interpretable trajectories.

We showcase the potential of the proposed approach in the contexts of two specific clinical problems with different scopes and challenges: (1) nonstandardized stress echocardiography and (2) labour monitoring and decision making.

In the first application, the proposed approach proved to help in the identification of normal and abnormal patterns in cardiac response to stress and in the understanding of the underlying pathophysiological mechanisms, in a context of nonstandardized longitudinal data collection involving heterogeneous data streams. In the second application, we showed how the proposed approach could be used as the central concept of a personalized labour monitoring and decision support system, outperforming the current reference labour monitoring and decision support tool.

Overall, we believe that this thesis validates unsupervised dimensionality reduction as a promising approach to the analysis of complex and nonstandardized clinical longitudinal data.

Resumen

El seguimiento de la evolución de un paciente tomando las mismas medidas en diferentes instantes temporales (datos longitudinales) es un componente crucial en la dinámica de los cuidados médicos, ya que permite tomar decisiones correctas en el momento idóneo para prevenir eventos adversos. Es entonces importante que los médicos tengan a su disposicion herramientas para analizar datos de carácter longitudinal. Sin embargo, la mayoría de las herramientas que actualmente existen tienen una aplicabilidad limitada si los datos (1) no están suficientemente estandarizados o (2) son muy heterogéneos (eg: imágenes, señales, variables continuas y categóricas) y/o tienen una alta dimensionalidad. Estas limitaciones son tremendamente relevantes, ya que ambos casos son prevalentes en la practica clínica habitual.

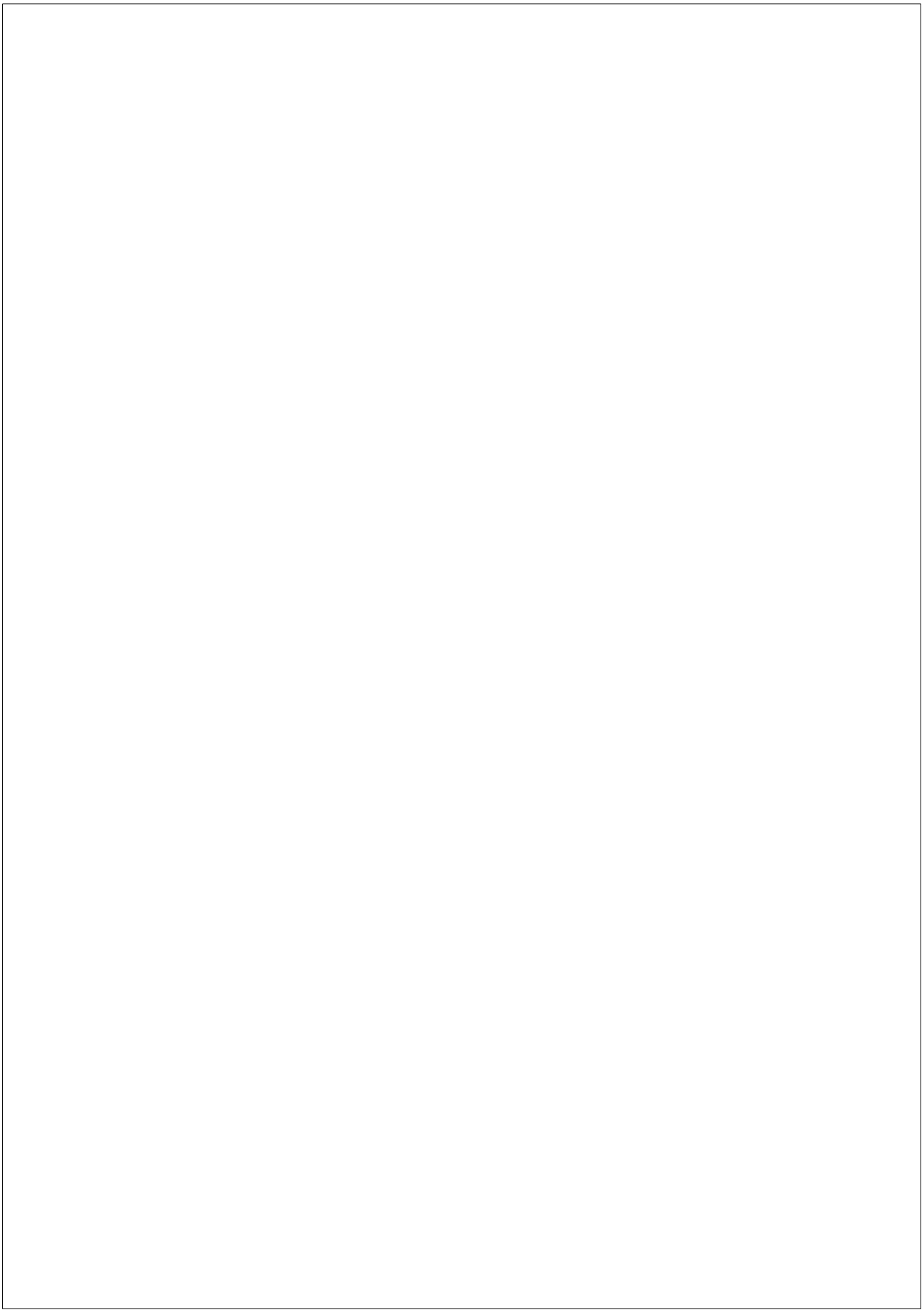
El objetivo de esta tesis es el desarrollo de herramientas que facilitan la integración e interpretación de datos clínicos longitudinales que son complejos y no están estandarizados. Específicamente, exploramos enfoques basados en la reducción de dimensionalidad no supervisada, que permite integrar datos longitudinales complejos y su representación como una trayectoria de baja dimensión que es clínicamente interpretable.

Mostramos el potencial del enfoque propuesto en el contexto de dos problemas clínicos en diferentes ámbitos y con diferentes desafíos: (1) ecocardiografía de estrés no estandarizada y (2) monitoreo de parto y toma de decisiones.

En la primera aplicación, el enfoque propuesto ha mostrado ser de ayuda en la identificación de patrones normales y anormales en la respuesta cardiaca al estrés y en entender los mecanismos patofisiologicos subyacentes, en el contexto de una adquisición de datos longitudinales no estandarizados que contiene un flujo de datos heterogéneo. En la segunda aplicación, mostramos como el enfoque propuesto puede ser el concepto central de un sistema de monitoreo del parto y soporte a la decisión personalizado, superando el sistema actual de referencia.

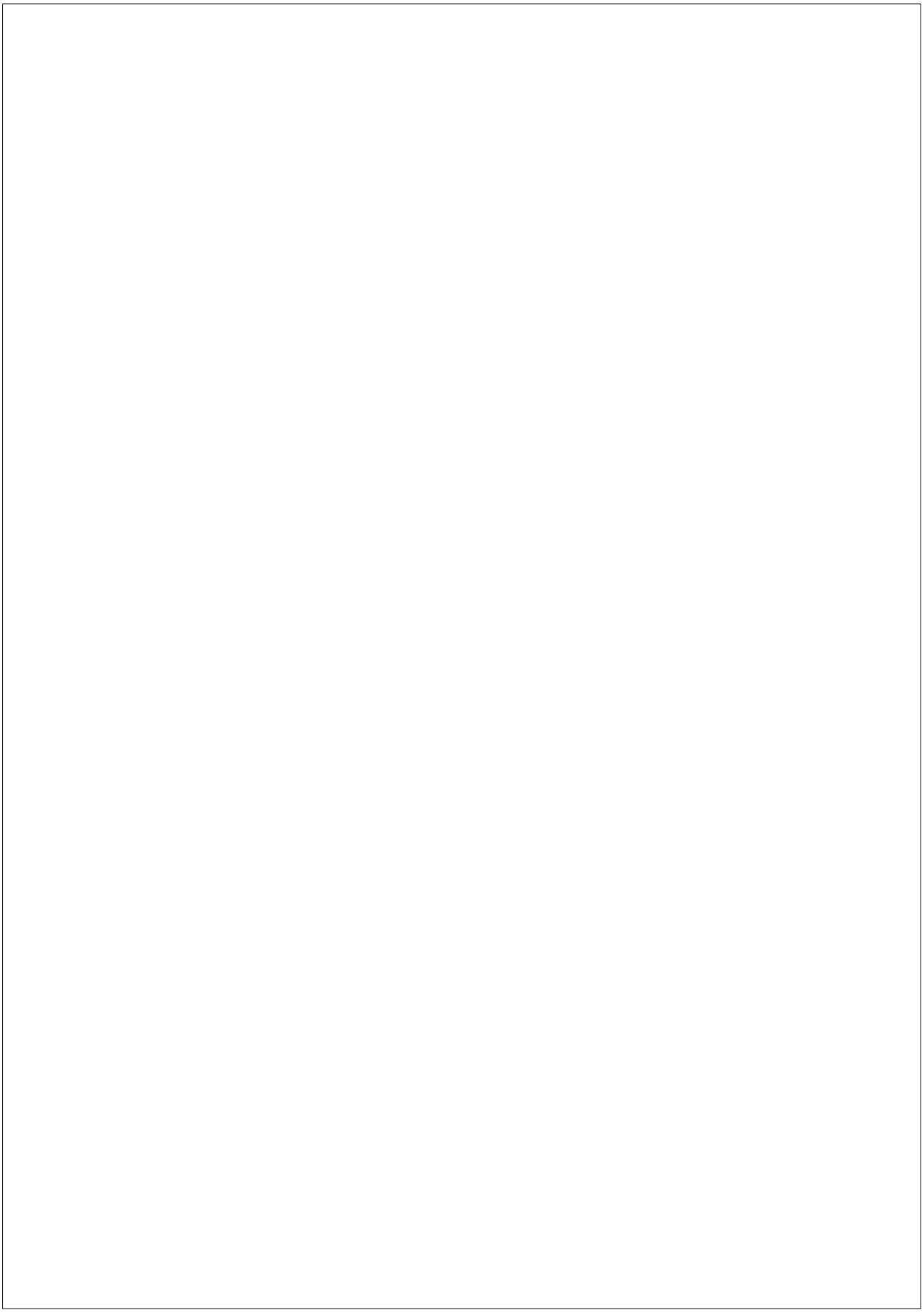
En conclusión, creemos que esta tesis muestra que la reducción

de dimensión no supervisada es un prometedor enfoque para analizar datos clínicos longitudinales complejos y no estandarizados.

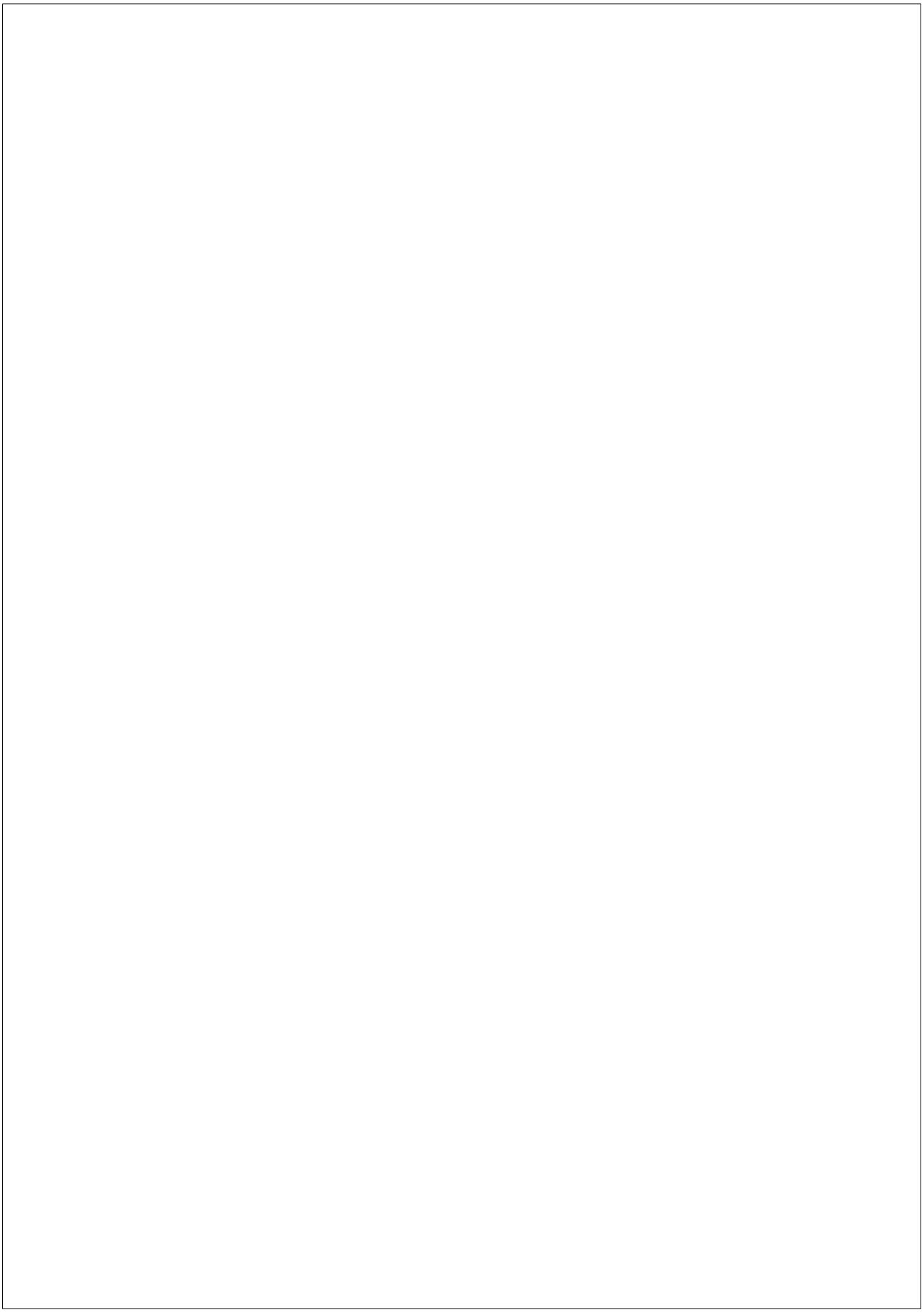


Contents

List of Figures



List of Tables



Acronyms

ADP adenosine diphosphate.

ANT1 Adenine Nucleotide Translocator-1.

API application programming interface.

ATP adenosine triphosphate.

AUC area under the receiver operating characteristic curve.

BO bad outcome.

CS caesarean section.

DMI Doppler myocardial velocity imaging.

DNS domain name system.

DSS decision support system.

DTW dynamic time warping.

ECG electrocardiogram.

FNR false negative rate.

FPR false positive rate.

GEP generalized eigenvalue problem.

HFPEF heart failure with preserved ejection fraction.

HR heart rate.

HTTPS hyper text transfer protocol secure.

JWT JSON web token.

MDS multidimensional scaling.

MKL multiple kernel learning.

MKR multiscale kernel regression.

ML machine learning.

MPI message passing interface.

NPV negative predictive value.

OpenMP open multi-processing.

PPV positive predictive value.

SDP semidefinite programming problem.

SE sensitivity.

SELMA Simplified, Effective, Labour Monitoring-to-Action.

SP specificity.

URL uniform resource locator.

WHO World Health Organization.

Chapter 1

INTRODUCTION

1.1. Context and Motivation

1.1.1. Overall

Keeping up with patients' health status by performing repeated measurements over time is a common practice in clinical care that can be crucial for preventing adverse outcome, by allowing the clinician to make timely decisions in that sense. A diagram depicting the typical decision-making process in clinical practice is shown in Figure 1.1. A first set of data are acquired, which are descriptors of the patient's state. The clinician then has the task of integrating the data and comparing them with those of previous patients, in order to position the current patient in the disease spectrum, while taking into account the uncertainty and reliability associated with the available information. After all these considerations, the clinician can decide that (1) the available information is insufficient and request the acquisition of complementary data, (2) an intervention is necessary or (3) no further action is currently necessary – but paying careful attention to patient evolution. In all cases, the decision most often implies posterior re-acquisitions of data to monitor natural evolution or response to intervention, and the loop in Figure 1.1 is restarted.

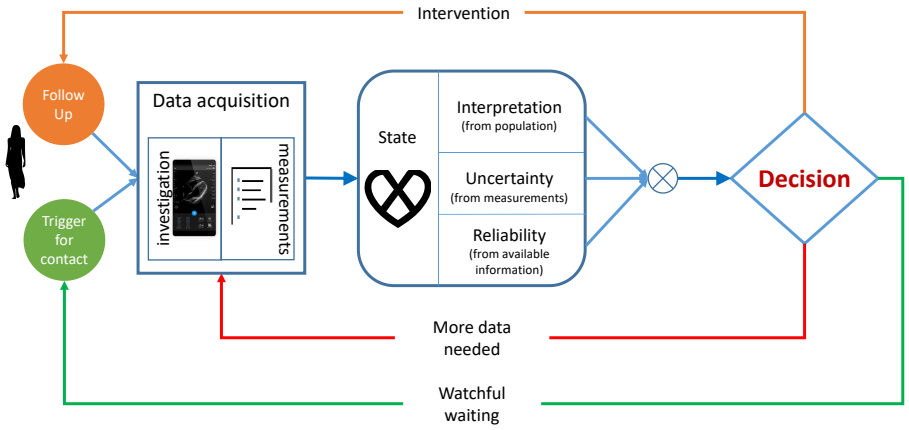


Figure 1.1: Decision making in clinical practice.

In clinical trials, every aspect of this chain of events, from inclusion to data acquisition, treatment and follow-up is strictly protocolized and standardized. On the contrary, clinical routine is highly nonstandardized: patients are heterogeneous in demographics and medical backgrounds, and come in with variable symptoms; the way each patient is handled upon arrival (e.g. hospitalization, data acquisition) is highly dependent on local practice and resources; the decision of the clinician in terms of treatment is greatly influenced by experience, and timings of follow-ups are also not standardized. The standardized nature of clinical trials largely simplifies the data analysis process in the sense of answering a specific research question. Trying to answer the same question becomes a much more complicated task in settings that are closer to clinical routine. A diagram highlighting the differences in standardization between clinical trials and clinical routine is depicted in Figure 1.2.

Nowadays, longitudinal data can be very rich, high-dimensional, and comprise very heterogeneous types of temporally changing data (continuous or categorical variables, curves, images, etc.). Integrating

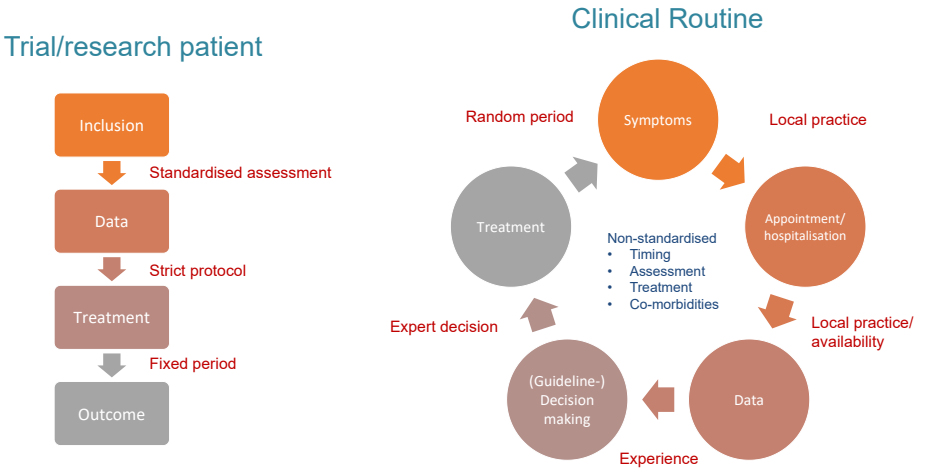


Figure 1.2: Standardized nature of clinical trials versus nonstandardized nature of clinical routine.

and making sense of all this information can already represent a challenge for the clinician. Indeed, because of this difficulty, a common approach is to extract established simplified measurements from the complex original data and base the analysis on the former. However, by doing so, the clinician is likely losing valuable information. If, on top of high-dimensional and/or heterogeneous, data are nonstandardized, the task becomes all the more challenging. Traditional longitudinal analysis methods are not prepared to handle such complex data, and clinicians lack the tools to assist them in taking proper advantage of their potential.

In this thesis, we aim to develop alternative tools that facilitate the integration and interpretation of this type of data. Specifically, we explore the potential of tools centered on unsupervised multiview dimensionality reduction, which allow us to operate on lower-dimensional yet interpretable representations of the data (Figure 1.3). The objective of this thesis is to describe and showcase the potential of this type of approach while addressing two specific clinical problems with different

scopes and challenges: (1) non-standardized stress echocardiography and (2) labour monitoring and decision making.

Herein, the two clinical problems to be addressed are briefly introduced.

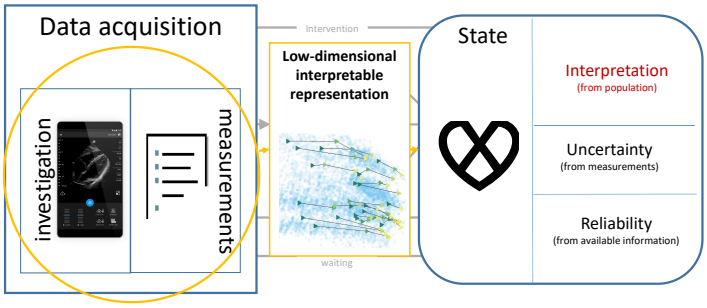


Figure 1.3: Scope of this thesis’ work in the context of the flowchart in Figure 1.1 – facilitating integration and interpretation of data through the development of tools that build upon lower-dimensional yet interpretable representations.

1.1.2. Application I: Nonstandardized stress echocardiography

Stress echocardiography plays an important role in the screening and study of cardiovascular disease. During a stress test, the heart is put under some type of stress and is expected to develop stress-specific adaptation mechanisms in the cardiac cycle that correspond to a healthy response. One test typically spans resting, build up, peak stress, and recovery periods, and cardiac imaging and signal data are acquired throughout in order to assess whether response is healthy or abnormal. The most common stress inducers in clinical practice are exercise (e.g. using a treadmill) or pharmacological agents such as dobutamine [Voigt, 2003, Davidavicius et al., 2003].

In classical exercise and pharmacological protocols, stress levels are relatively easy to quantify and control, which is very convenient in the sense that it makes them highly standardizeable: a few stress levels can be chosen beforehand, and the same protocol can be used to test all subjects. This allows comparisons to be made in a rather direct and quantitative way, based only on data corresponding to a few representative heartbeats out of the whole test. However, both exercise and pharmacological protocols involve high costs, as they require highly-skilled staff and expensive equipment, and are rather time-consuming, which ultimately translates into limited applicability.

Aiming for large-scale applicability would imply simpler, relatively inexpensive, practical and low-risk protocols. For this reason, efforts have been directed towards exploring less typical protocols as potential alternatives, such as the handgrip or cold pressor tests [Helfant et al., 1971, Velasco et al., 1997]. There is, however, one inconvenience when transitioning to these types of protocols – they require a change of paradigm regarding the way stress echocardiography data is analysed, as they are not standardizeable, i.e., quantifying and controlling stress levels is no longer straightforward. This calls for an approach that relies on the identification of patterns and trends within the spectrum of stress levels that make up the full acquisition. The clinician is then left with the task of identifying such patterns and trends while integrating information coming from multiple heterogeneous data channels and spanning dozens of cardiac cycles.

In the context of this application, the goal of this thesis is the development of tools to assist the clinician in the processes of integration, visualization, and interpretation of this type of data.

1.1.3. Application II: Labour monitoring and decision making

The majority of pregnancy-related deaths and morbidities have origin around the time of childbirth [Oladapo et al., 2015]. Quality of care during labour is thus of critical importance, and a continu-

ous monitoring process is essential to allow timely decision-making in order to prevent adverse outcomes. However, there is still much uncertainty regarding which is the optimal approach to labour monitoring and decision-making [Robson et al., 2015]. Decades ago, the World Health Organization (WHO) introduced the partograph in an effort to standardize practice. In the partograph, the healthcare provider writes down and tracks the evolution of multiple fetal and maternal measurements over time (in nonstandardized intervals) and evaluates whether they are progressing as expected or not, in which case the reinforcement of monitoring or actual intervention might be necessary. The pace of cervical dilatation is given a central role in the partograph as an indicator of normality or abnormality in labour progress, building upon Friedman's "1 cm/h rule", which sets a lower-limit for a normal dilatation progress at 1 cm/h as of the onset of 4 cm [Friedman, 1954]. However, multiple studies have demonstrated that the concept of "normal" spontaneous labour can vary depending on the particular characteristics of the pregnant woman (e.g. demographics, previous pregnancy history, and others), and the partograph's one-fits-all approach to the diagnosis of abnormal labour progress has received much criticism [Souza et al., 2015]. Furthermore, there is lacking evidence on the positive impact of its use [Souza et al., 2015]. These are some of the reasons, among others, as to why the partograph has effectively failed to be fully incorporated in clinical practice and thus to achieve its ultimate objective to standardize it. Currently, not only is practice (and outcome) very heterogeneous, but there is a global increase in rates of interventions, such as caesarean sections, which bear risks for the mother and child – a concerning trend that is all the more worrying in settings where resources are limited and intervention-associated risks are amplified [Betrán et al., 2018, Boatin et al., 2018].

In this context, the WHO has identified the need for the development of novel labour management tools that identify abnormal labour progress, risk of adverse outcome and need for intervention, in a personalized and evidence-based way [Oladapo et al., 2015, Souza

et al., 2015].

Within the scope of this application, the objective of this thesis is the formulation of a methodological pipeline designed to address such need, its evaluation, and its integration into a prototype of a user-oriented tool designed to be deployed in a real-world clinical setting.

1.2. Proposed approach

While the nature, scope, and challenges of the addressed clinical problems are very different, there is a common need for simplifying the integration and understanding of complex longitudinal data, and a lack of available tools to respond to it. In this thesis, we address this methodological gap that affects these and many other clinical problems.

Analysing simplified, lower-dimensional latent representations is a common strategy to cope with high-dimensional and complex data. Nonetheless, when it comes to longitudinal data, it is a rather under-explored practice. In this thesis, we explore the potential of this approach to the analysis of complex longitudinal data.

In the latent space, complex longitudinal data can be visualized as low-dimensional yet clinically interpretable trajectories. Our hypothesis is that moving the analysis process to this simplified space can help in the identification of normal and abnormal evolution patterns and underlying causes. The way in which the lower-dimensional space information is processed and used, posterior to the dimensionality reduction step, is more scope-dependent and, thus, approached in an application-specific way, which is described in detail in the corresponding chapters.

The analysis tools developed in this thesis build upon a specific dimensionality reduction algorithm: unsupervised multiple kernel learning (MKL) [Lin YY, 2011, Sanchez-Martinez et al., 2017]. MKL allows the integration of heterogeneous features by first mapping all

of them into a unified representation – similarity matrices. After this step, data can be easily combined. The similarity information from the different features is then merged in order to learn a projection model to a lower-dimensional space where distance relations are dictated by similarity relations in the input feature space.

The choice for this algorithm was tied with our objective of developing an approach that would (1) be able to easily accommodate and integrate numerous and very heterogeneous views of data (multiview data), making it flexible enough to be applicable to any clinical problem, (2) have the latent representation of the data be learned in an unsupervised way, thus free of constraints imposed by labellings that are often inaccurate, biased, or misrepresentative of the variability of the target variable and (3) have it instead be dictated by patient similarity, which allows for an interpretation process that is very close to clinical practice.

An illustration of the approach is depicted in Figure 1.4.

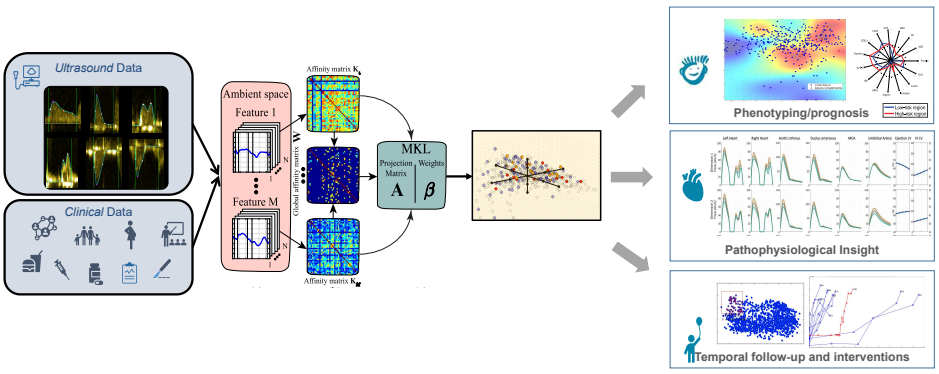


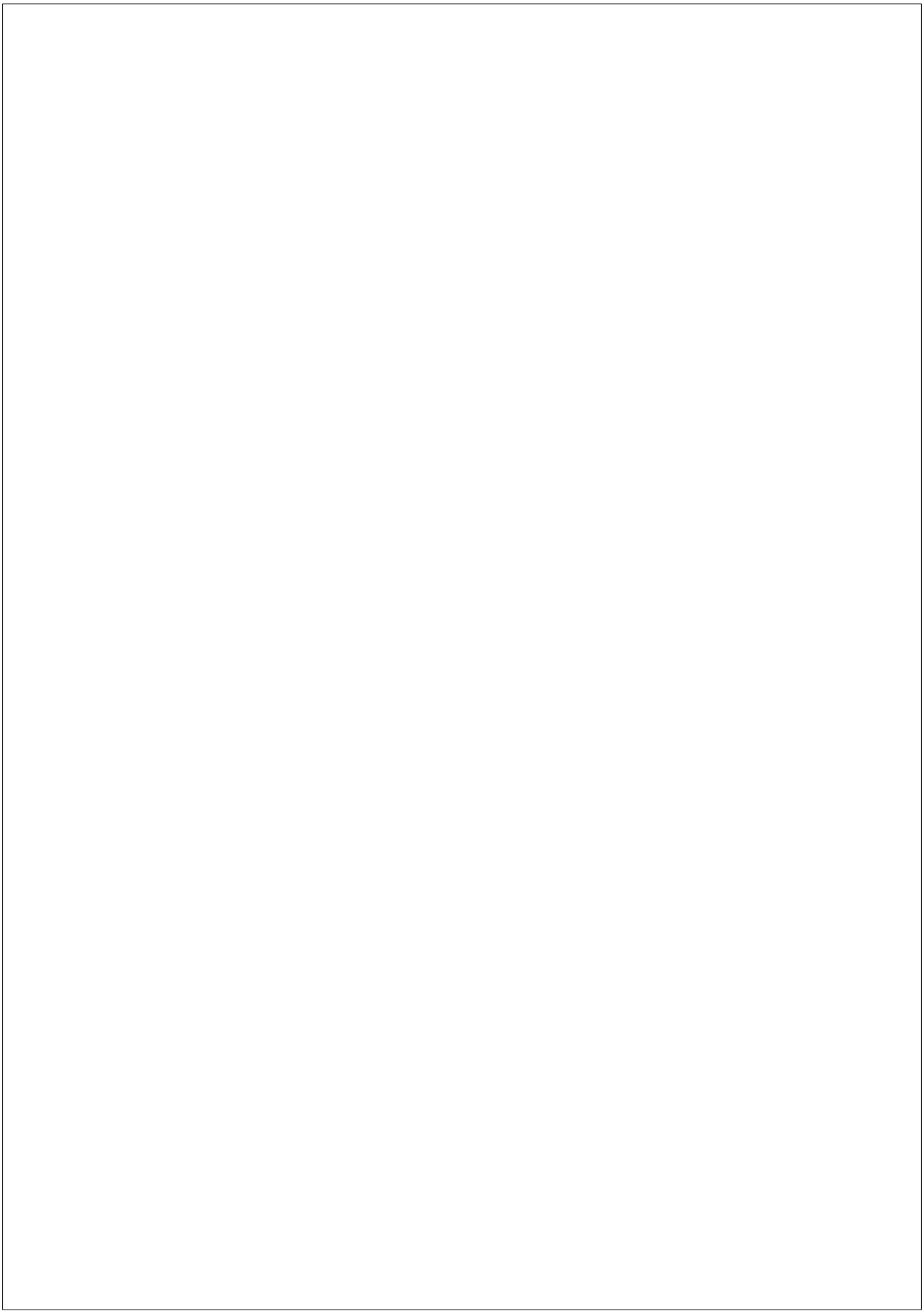
Figure 1.4: Illustration of the proposed approach.

1.3. Thesis outline

This thesis is composed of three main, self-contained chapters, that are presented in the form of research/implementation papers. One of

the chapters refers to the work developed in the aim of *application I*, while that developed in the aim of *application II* is split into two chapters, one dedicated to the formulation and evaluation of the proposed methodology, and another dedicated to the description of its materialization into a functional prototype of a real-world clinical setting tool. The remainder of this manuscript is thus organized as follows:

- **Chapter 2.** A methodological framework is developed for the analysis of nonstandardized stress echocardiography sequences. For illustration and evaluation purposes, the proposed framework is applied to the comparative analysis of handgrip test sequences of a cohort composed of healthy controls and Adenine Nucleotide Translocator-1 (ANT1)-associated mutation patients.
- **Chapter 3.** A methodological framework is developed for labour monitoring and decision support. For illustration and evaluation purposes, the proposed framework is applied in the prediction of intervention and outcome in WHO's Simplified, Effective, Labour Monitoring-to-Action (SELMA) dataset.
- **Chapter 4.** Description of the integration of the methodological framework developed in *Chapter 3* in a functional prototype of a user-oriented tool designed to be deployed in a real-world clinical setting.
- **Conclusion.** Summary of the main contributions of the presented work, limitations, and future directions.



Chapter 2

ANALYSIS OF NONSTANDARDIZED STRESS ECHOCARDIOGRAPHY SEQUENCES USING MULTIVIEW DIMENSIONALITY REDUCTION

This chapter is adapted from: M. Nogueira, M. De Craene, S. Sanchez-Martinez, D. Chowdhury, B. Bijnens, G. Piella. Analysis of nonstandardized stress echocardiography sequences using multiview dimensionality reduction. *Medical Image Analysis*, 60:101594.

2.1. Introduction

2.1.1. Clinical Context and Motivation

Stress echocardiography can unveil early-stage cardiovascular-pathology signatures that are not expressed at baseline condition, thus being a valuable tool for screening purposes. Current stress echocardiography protocols, based on exercise or pharmacological stress [Voigt, 2003, Davidavicius et al., 2003], are standardized, meaning that the control of the stress levels over the test is very rigorous (based on dose, heart rate, time, etc.), allowing the evaluation of response to stress to be performed based on the comparison of measurements collected at a few discrete timepoints (corresponding to very precise stress levels). However, this standardization comes at the cost of cumbersome protocols, being time-consuming as well as requiring highly-trained staff and specialized equipment. All this translates into high costs, which limit the application of current protocols to a fairly lesser extent than desired, and thus making them unsuited for large-scale screening purposes. Moreover, by getting data at pre-determined intervals and timings, one might be missing pertinent information, as the disregarded dynamic data potentially contain additional valuable information concerning the patient's physiological state.

Other less standard forms of stress, such as the cold pressor test [Velasco et al., 1997] and handgrip exercise [Strauss et al., 2013, Kivowitz et al., 1971, Helfant et al., 1971], were already reported to trigger cardiovascular responses that could unmask differential responses to stress by healthy and pathological patients. These protocols are cheap and practical, come with low risks to the patient, and involve little patient motion, making imaging an easier task. As such, they hold great potential for screening, overcoming the main limitations of current protocols. Besides their potential for screening, they also represent an alternative for patients that are physically unable to undergo a classical exercise test. However, there is one main drawback: the level of exercise is hard to quantify and control, and the timings

and magnitudes of events are unpredictable; in other words, they are nonstandardizable. In practice, this implies continuously analyzing the complete acquisition, and focusing on trends/patterns of response rather than on a discrete set of values. The analysis, of these long, dynamic, heterogeneous sequences, which also implies integration of multiple features, is not trivial, and clinicians lack tools to assist them in this task. On the other hand, this type of analysis may be advantageous: by allowing the identification of variations in exercise performance throughout the dynamic range (versus at discrete points in time), it may be more informative of the patients physiological state, and thus have a higher predictive value of adverse outcomes.

Currently, machine learning is being established as one of the preferred tools for the analysis of patterns in functional and high-dimensional data, and has become remarkably popular within the biomedical field. It has already been applied to the study of cardiac response to stress, based on multiple heterogeneous descriptors, such as the velocity profiles of different myocardial segments and timings of key events in the cardiac cycle [Sanchez-Martinez et al., 2017, Sanchez-Martinez et al., 2018]. However, to the best of our knowledge, it has not yet been used to explore nonstandardized continuous echocardiographic recordings. In this paper, we propose an analysis framework that explicitly addresses the practical challenges this kind of sequences pose, and illustrate its potential in a specific group of cardiac patients.

2.1.2. Technical Context

In biomedical research, there is an emergent need for machine learning algorithms able to learn from multiple concurrent data sources (e.g. imaging, signal, patient metadata). This type of learning is commonly referred to as multiview learning [Xu et al., 2013]. In the cardiac domain, both supervised and unsupervised multiview learning algorithms have been recently applied in the analysis of cardiac motion patterns for numerous applications, e.g. in the identification of

dilated cardiomyopathy [Puyol-Antón et al., 2019], in cardiac resynchronisation therapy response prediction [Peressutti et al., 2017] or in the study of heart failure with preserved ejection fraction (HFPEF) [Sanchez-Martinez et al., 2017, Sanchez-Martinez et al., 2018].

In this work, we integrate information coming from multiple heterogeneous features (i.e., heart rate and velocity traces from echocardiographic images) to evaluate patterns of response to stress. Since nonstandardized sequences typically last 60-120 cardiac cycles (equivalent to thousands of images), we propose unsupervised multiview dimensionality reduction to obtain a compact representation of the patterns of response over time. This low-dimensional representation can be used to obtain the principal modes of variation – which describe how the features change – and the temporal trajectories – which encode the timings and intensity of such changes.

Unsupervised multiview dimensionality reduction is an active field of research, including canonical correlation analysis [Hotelling, 1936], partial least squares [Wold, 1985], multiple kernel learning (MKL) [Lin YY, 2011] or multi-modal autoencoders [Li et al., 2018] as some of the most popular algorithms. Our choice for MKL was based on (1) its ability to address inherent nonlinearities of the data and any number of desired input features, without strong assumptions on their correlations, and (2) its good performance in similar applications, while providing a fairly simpler, very flexible, potentially more intuitive/interpretable framework than other types of machine learning.

Once a low-dimensional embedding is estimated, the main modes of variation in the data can be reconstructed using multiscale kernel regression (MKR) [Bermanis et al., 2013, Duchateau et al., 2013]. A combined analysis using MKL and MKR was successfully explored before by [Sanchez-Martinez et al., 2017, Sanchez-Martinez et al., 2018] to characterize functional responses to semi-supine bicycle exercise of controls and patients with HFPEF, based on left-ventricular velocity patterns. This work dealt, however, with only two-timepoint (rest/stress) information for each patient, acquired during a standard-

ized exercise stress test.

We propose a technical framework that extends this analysis to the challenging context of nonstandardized stress echo datasets.

2.1.3. Proposed Approach

Our framework uses MKL to project heterogeneous data collected at each cardiac cycle throughout the stress test onto a low-dimensional space where the main variations in the data are encoded. In this space, the response to stress of each subject can be seen as a trajectory and, based on the similarity among trajectories, subjects can be grouped in clusters that reflect differential patterns of response. The physiological interpretation of the results is decoded through MKR, which allows reconstructing the input signals along any path over the low-dimensional output space.

A preliminary version of the framework was previously proposed [Nogueira et al., 2017]. The present paper extends the work in several aspects: we test the framework against a real dataset including healthy and pathological cases, whereas previously the cases had been generated synthetically; we explore other physiological features, using velocity traces at the basal septum of the left ventricle instead of the global longitudinal strain; we reformulate the clustering analysis in the trajectory space by exploring a more sophisticated way of computing distances among trajectories, involving dynamic time warping (DTW) [Bemdt and Clifford, 1994]. In addition, we enrich the analysis by exploring and interpreting the spatial configurations of the distributions of the control and diseased population samples in the output space.

2.1.4. ANT1 mutation

To illustrate the framework, we apply it to the discriminative analysis between the dynamics of response to stress in patients with Adenine Nucleotide Translocator-1 (ANT1) deficiency (due to a mu-

tation in an encoding gene) and controls, during handgrip exercise challenges. In patients with ANT1 mutation there is a lack of adenine nucleotide transferase, which converts ADP to ATP. The decreased availability of ATP to the muscles causes lactic acidosis. These patients present with shortness of breath with exercise at a very young age. Within the scope of this paper, they can be considered as extreme cases of HFPEF.

2.2. Methods

A diagram illustrating the main blocks of the framework is depicted in Figure 2.1. The first block corresponds to the automated processing and extraction of features from the sequence data (Section 2.2.2). The second block refers to the application of MKL to obtain a low-dimensional representation of the data (Section 2.2.3). Finally, the third block corresponds to the analysis of this low-dimensional representation, focusing on the discrimination between groups of response and the understanding of the underlying pathophysiological mechanisms (Section 2.2.4).

2.2.1. Data

This study includes 15 subjects, 10 controls (average age 24 ± 14 years) and 5 ANT1 mutation patients (average age 21 ± 7 years). The echocardiographic acquisitions were performed using a Vivid Q system (GE Healthcare). For each subject, a Doppler myocardial velocity imaging (DMI) sequence of the apical 4-chamber view was acquired (average sampling rate 115 ± 43 Hz) during handgrip exercise. All sequences comprise the start of exercise, a phase of sustained exercise and recovery (average heart rate 92 ± 18 bpm for controls; 118 ± 23 bpm for ANT1 patients). The durations of each phase vary across subjects. When the 15 subjects are considered, our dataset amounts to a total of 1377 cardiac cycles (average sequence length 92 ± 26 cardiac cycles).

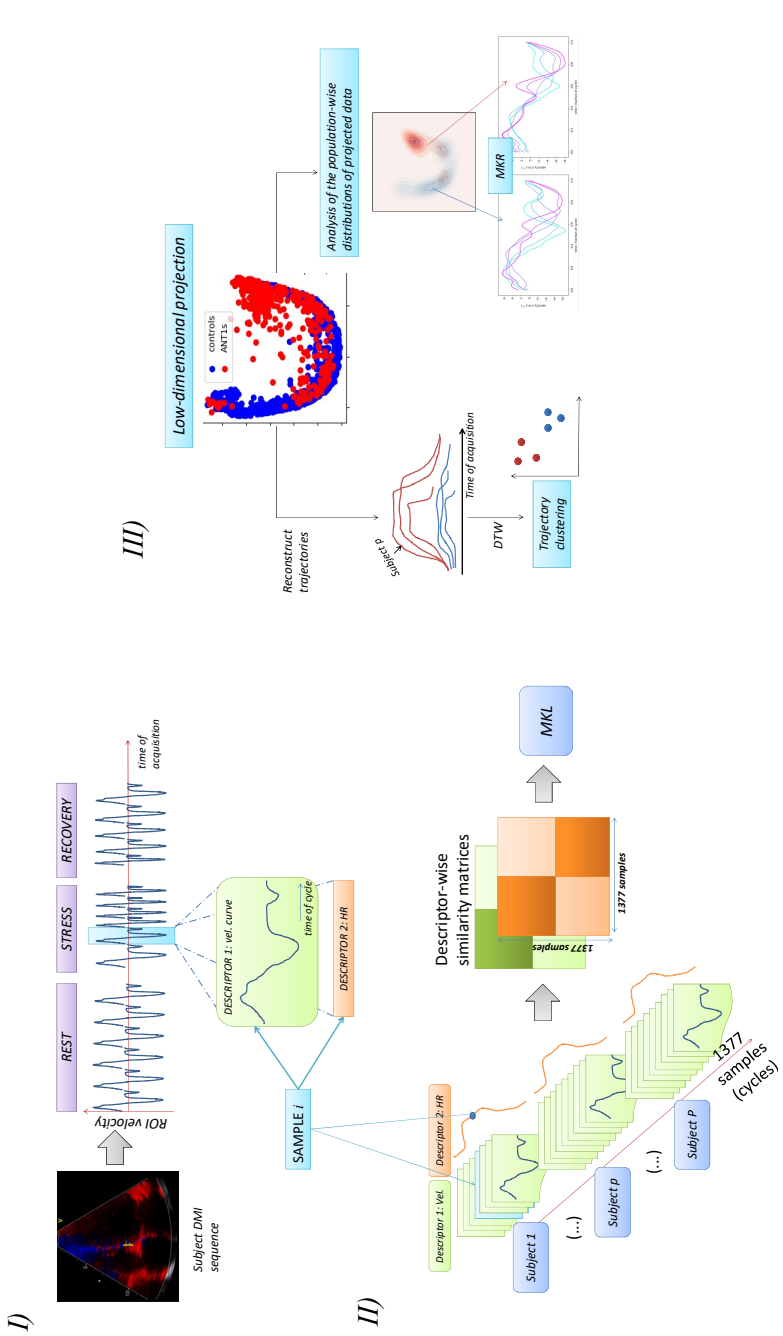
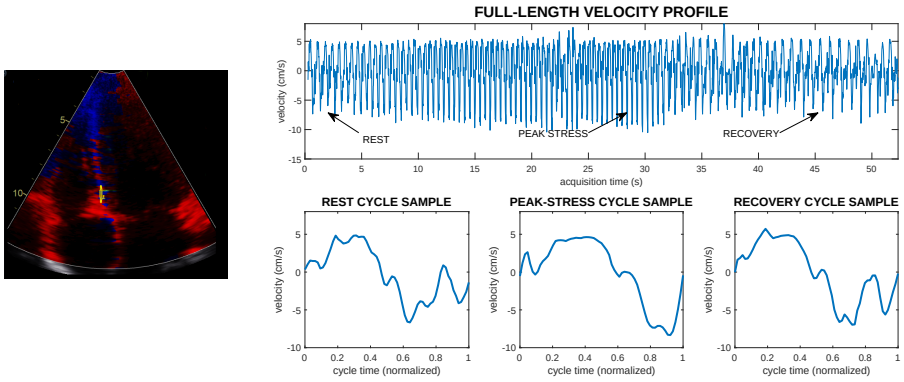


Figure 2.1: Main stages of the proposed framework: *i)* automated cycle-wise feature extraction; *ii)* multiview dimensionality reduction to project stress echo sequences onto a low-dimensional space; *iii)* physiological interpretation of the output-space sample distributions and cluster analysis in the trajectory space. DMI=Doppler Myocardial Velocity Imaging. HR=Heart Rate. MKL=Multiple Kernel Learning. DTW=Dynamic Time Warping. MKR=Multiscale Kernel Regression.



(a) (b)

Figure 2.2: Extraction of velocity sequence data. (a) Example of a frame from a DTI sequence from an ANT1 patient. The yellow circle over the basal septum is the region of interest used to monitor the velocity over the whole sequence. (b) Top: example of a full-length velocity trace. Bottom: isolated rest (left), peak-stress (middle) and recovery (right) cycles, extracted from the corresponding annotated regions in the top plot.

2.2.2. Feature Extraction

In our dataset, we have an average of about 70 DMI frames per cardiac cycle. As such, 1377 cycles contain a considerably large amount of data, calling for integration and simplification. The first simplification comes with feature extraction, i.e., collecting relevant descriptors of cardiac function throughout the acquisition, while ensuring their robustness to noise and artifacts in the data (e.g. due to breathing or transducer motion). Features should be easy to obtain in clinical practice and, ideally, in an automated manner (manually processing these many cardiac cycles would be impractical).

We selected the left-ventricular basal-septum velocity profile and heart rate (HR) as the features of interest to monitor during the stress protocol. These were automatically extracted with the aid

of the ECG as temporal reference: the full-length velocity traces were extracted from the DMI sequences using a commercial software (EchoPAC, v.113, GE Healthcare), by manually placing a region of interest (default dimensions) at the basal septal region (see Figure 2.2a; the actual trace is computed and exported through the software). The cycle-wise traces were obtained by slicing the full-length profiles at the R-peak positions of the simultaneously acquired ECG and the HR was obtained from the timings of the R peaks (whole process illustrated in stage *I* of Figure 2.1). Examples of a full-length velocity trace and sliced cycles (time-normalized for cycle duration as explained in 2.2.3) are featured in Figure 2.2b.

Finally, we fed a set of 1377 multiview samples (corresponding to all cardiac cycles of all 15 subjects) to the MKL algorithm, describing each cardiac cycle of each patient by a velocity curve and a HR value (see stage *II* of Figure 2.1).

2.2.3. Computation of the low-dimensional space using MKL

Given a high-dimensional dataset with N samples $X = \{x_i \in \mathbb{R}^d\}_{i=1}^N$, graph embedding aims at finding a low-dimensional projection $Y = \{y_i \in \mathbb{R}^k\}_{i=1}^N, k < d$, that preserves the main topology and variability of the data while removing noisy contributions. To achieve this, a similarity matrix W defined over the pairs of input samples is used to weight the optimization problem which, under appropriate constraints, can be generically expressed as

$$\min_Y \sum_{ij} \|y_i - y_j\|^2 W_{ij} \quad . \tag{2.1}$$

In this way, to minimize the product $\|y_i - y_j\|^2 W_{ij}$, close samples in the input space (high W_{ij}) are enforced to remain close in the output space (small $\|y_i - y_j\|$), while distant samples have little or no influence on each other's optimal projection.

Based on this graph embedding framework, Lin et al. [Lin YY, 2011] generalized the concept of MKL, originally formulated within the support vector machine framework [Bach et al., 2004, Hearst et al., 1998] for classification/regression, to (supervised and unsupervised) dimensionality reduction. By combining multiple kernels, each one based on a specific data descriptor, MKL fuses heterogeneous information and provides the contribution of each feature to the low-dimensional output representation. The unsupervised formulation, adopted in this work, can be summarized as follows.

Let the input dataset, composed of N samples with M descriptors each, be defined as $X = \{x_i\}_{i=1}^{N}, x_i = \{x_i^m \in \mathbb{R}^{d_m}\}_{m=1}^{M}$ where x_i^m represents the descriptor m associated with sample i and of dimensionality d_m. The projection of a sample is parametrized by a projection matrix $A \in \mathbb{R}^{N \times k}$ (where k refers to the selected dimensionality of the output space, $k \in [1, N-1]$) and a vector $\beta \in \mathbb{R}^M$ that determines the normalized weight of each feature in the mapping. A unified mapping based on heterogeneous descriptors is made possible as A and β operate on kernelized data rather than on their raw content. For each feature, a kernel matrix K_m is defined, encoding the similarities over the pairs of samples, based on kernel functions k_m, i.e.,

$$K_m \in \mathbb{R}^{N \times N} \quad \text{with} \quad K_m(i,j) = k_m(x_i^m, x_j^m) \quad . \tag{2.2}$$

In this work, k_m is a Gaussian kernel (with Euclidean distance) whose bandwidth σ_m is computed as the average of the pairwise Euclidean distances between each descriptor x_i^m and its K nearest neighbors $\{x_{ij}^m\}_{j=1}^{K}$[Sanchez-Martinez et al., 2017]. The input descriptors we consider here are the HR (i.e. $x_i^1 \in \mathbb{R}$) and the longitudinal velocity values along each cycle. As the dimension of the latter varies over cycles, all cycles were resampled along the temporal axis so that $x_i^2 \in \mathbb{R}^{d_2}$ (we set $d_2 = 65$).

Based on $\{K_m\}_{m=1}^{M}$, a set of sample-wise matrices $\{\mathbb{K}^i\}_{i=1}^{N}$ is defined. Each $\mathbb{K}^i$ encodes the similarity of sample i to the other samples taking into account the different descriptors. In practice, $\mathbb{K}^i$ is built from stacking the i^{th} columns of all kernel matrices $\{K_m\}_{m=1}^{M}$.

Formally, the projection of sample i is expressed as

$$y_i = A^T \mathbb{K}^i \beta \quad . \tag{2.3}$$

Plugging (2.3) into (2.1), the optimization problem becomes

$$\min_{A,\beta} \sum_{i,j} \|A^T \mathbb{K}^i \beta - A^T \mathbb{K}^j \beta\|^2 \mathbb{W}_{ij} \tag{2.4}$$

$$\text{s.t.} \sum_i \|A^T \mathbb{K}^i \beta\|^2 \mathbb{D}_{ii} = 1, \tag{2.5}$$

$$\beta_m \geq 0, \sum_m \beta_m = 1 \tag{2.6}$$

where $\mathbb{W}$ is the multiview generalization of W in (2.1), a global affinity matrix computed by combining all the individual kernel matrices (in this paper we used $\mathbb{W} = \frac{1}{M} \sum_m K_m$, with kernel matrices $\{K_m\}_{m=1}^M$ being normalized across features prior to the summation through a variance-based method, described by [Sanchez-Martinez et al., 2017]). The constraint in (3.7), with $\mathbb{D}_{ii} = \sum_j \mathbb{W}_{ij}$, removes an arbitrary scaling factor in the output embedding.

Minimizers A^* and β^* are obtained by an iterative two-step optimization strategy [Lin YY, 2011]. At each iteration, A and β are alternately fixed to the value of last-step's solution and the problem is solved for the other. Iterations stop once a convergence criterion is met (e.g. maximum number of iterations or stable value of cost function). Solving (3.8) for A amounts to a generalized eigenvalue problem: the columns of the optimal A are the corresponding eigenvectors. Solving (3.8) for β, on the other hand, corresponds to a nonconvex quadratically constrained quadratic programming problem. To obtain a low-dimensional representation, one can choose the columns of A associated to the k lowest eigenvalues, yielding $A \in \mathbb{R}^{N \times k}$ and thus $y_i \in \mathbb{R}^k, \quad i = 1, ..., N$.

Once A and β have been learnt, the projections of the training samples can be computed using (2.3). Moreover, a new sample z can be mapped into the low-dimensional space by

$$y_z = A^T \mathbb{K}^z \beta \quad , \tag{2.7}$$

$\mathbb{K}^z \in \mathbb{R}^{N \times M}$ and $\mathbb{K}^z(n, m) = k_m(x_n, z)$.

Thus, the projection of new samples is determined by the similarities of their input-space features with those of the samples in the training set.

2.2.4. Discriminative analysis and physiological interpretation

In the low-dimensional space, the spatial distribution of the projected cycles is learned in an unsupervised way, solely based on their input-space similarities and not taking into account any label (i.e. control/ANT1) information. Our aim is to explore this simplified representation towards the identification of distinctive clusters of response by the two populations, and the unraveling of the pathophysiological mechanisms behind such differences.

We perform two levels of analysis (see stage *III* of Figure 2.1): one that is based on the overall spatial distribution of samples of each population in the output space (i.e., not distinguishing subjects), and another where we cluster the subjects based on the trajectories defined by their sequences in the output space.

2.2.4.1. Cycle-wise analysis: population signatures

To obtain the predominant patterns of response of each population, we *i)* draw a path passing through the regions of higher density of both healthy and diseased populations, and *ii)* sample the path at multiple points and adopt a multiscale adaptation of kernel regression (MKR) [Bermanis et al., 2013, Duchateau et al., 2013] to backproject them to input-space patterns. We hypothesize that analyzing the evolution of input features along this path will highlight discriminative characteristics of the diseased population.

Each such point q is backprojected based on an interpolation/regression from the known Y and X. A Gaussian kernel k is used to evaluate its similarity $k(q, y_i)$ with each $\{y_i\}_{i=1}^N \in Y$; its reconstruction in the space of feature m, here denoted as $f_m(q)$, is based on the known input-space representations $X_m = \{x_i^m\}_{i=1}^N$ and weighted by such similarities:

$$f_m(q) = \sum_i^N k(q, y_i) b_{mi} \tag{2.8}$$

where b_{mi} stands for the i^{th} column of matrix

$$B_m = \left(K + \frac{1}{\gamma_m} I\right)^{-1} X_m \tag{2.9}$$

with $K = [k(y_i, y_j)]$, γ_m a regularization weight and I the identity matrix. A multiscale approach is adopted where f_m is updated in an iterative coarse-to-fine process, with the kernel bandwidth halved at each step, from the maximum to the average output-space neighborhood size (details in [Duchateau et al., 2013]).

2.2.4.2. Sequence-wise analysis: subjects' trajectories

The idea behind the trajectory-based analysis is that the trajectories defined by the projected cycles of each subject (in temporal order) can be considered physiological descriptors of response to stress, and, as such, performing cluster analysis in the trajectory space can help us identify how all the subjects are organized in groups of response.

For each subject p, the trajectory defined by the projected data consists of a multidimensional $C_p \times k$ matrix, where C_p is the number of cycles of subject p's sequence. An element $(c, dim), \quad c = 1, ..., C_p, \quad dim = 1, ..., k$, tells us how the mode of variation associated with dimension dim is being expressed at cycle c. Intuitively, the whole trajectory matrix encodes a weighted combination of the k modes of variation at each cycle of the sequence. Our hypothesis

is that there will be differences in the trajectory matrices of the two populations, specific to the ANT1 pathology.

To cluster trajectories, as each subject's sequence has a different length, and different ratios of baseline/stress/recovery durations, standard distance metrics cannot be applied. For that reason, the DTW algorithm is used. This algorithm allows aligning two multidimensional time series by stretching sections in the temporal axis (one-to-many correspondence) in such way that some distance metric (Euclidean in our case) between the aligned time series is minimized [Bemdt and Clifford, 1994]. Prior to the DTW alignments, trajectories are slightly smoothed using total variation denoising [Rudin et al., 1992] (denoising weight $\lambda = 0.01$), to reduce noisy oscillations while preserving sharp transitions corresponding to state changes (rest-stress-recovery). Finally, a distance matrix is built from the pairwise distances and fed to a hierarchical clustering algorithm [Ward Jr., 1963], and the results are compared with the known labels.

2.3. Experiments and Results

2.3.1. Parameterization

Experiments were ran with several parameterizations. Alternatively to the standard iterative optimization process described in Section 2.2.3, having $\beta \in \mathbb{R}^2$ and $\sum_i \beta_i = 1, \quad \beta_i > 0$, we simply performed a grid search on a discrete set of vectors obeying $\beta = [\beta_1, 1 - \beta_1]^T, \quad 0 < \beta_1 < 1$, used them for initialization, and solved the corresponding generalized eigenvalue problem for the projection matrix A. In other words, we ran one single iteration of the standard optimization process for different initializations of the weight vector β.

Table 2.1 lists the parameterization corresponding to the results shown and discussed in this section. We denote by k_σ and k_{sparse} the number of neighbors used in the estimation of the kernel bandwidths and in a sparsing step of the global affinity matrix, respectively (refer

to [Sanchez-Martinez et al., 2017, Nogueira et al., 2017] for further details).

From our experiments, we found that the results presented a relatively low sensitivity to the values of β, k_{sparse} (except for very small values), and higher sensitivity to the value of k_σ. Lower kernel bandwidths mean higher sensitivity to variations in the data, and vice-versa. We heuristically tuned the value of k_σ having in sight a good trade-off between the spread and the spatial smoothness of the output-space data distribution.

For the MKR, we decided to use the first 6 dimensions of the projected data, since including further dimensions had little influence in the reconstructed modes (higher dimensions encode more noisy variability). The value of γ_m in (2.9) was tuned to minimize the average curve reconstruction error over 150 fixed samples (10 of each subject).

Table 2.1: Parameterization details. Feature weight vector defined as $\beta = [\beta_{HR}, \quad \beta_{velocity}]^T$.

Data	
N	1377
M	2
MKL	
k_σ	$0.05 \times N$
k_{sparse}	$0.25 \times N$
β	$[0.5, 0.5]^T$
MKR	
dimensionality k	6
γ_m	0.1

2.3.2. Population-wise analysis: representative signatures

We computed the low-dimensional representation of the data using MKL, with the parameterization in Table 2.1.

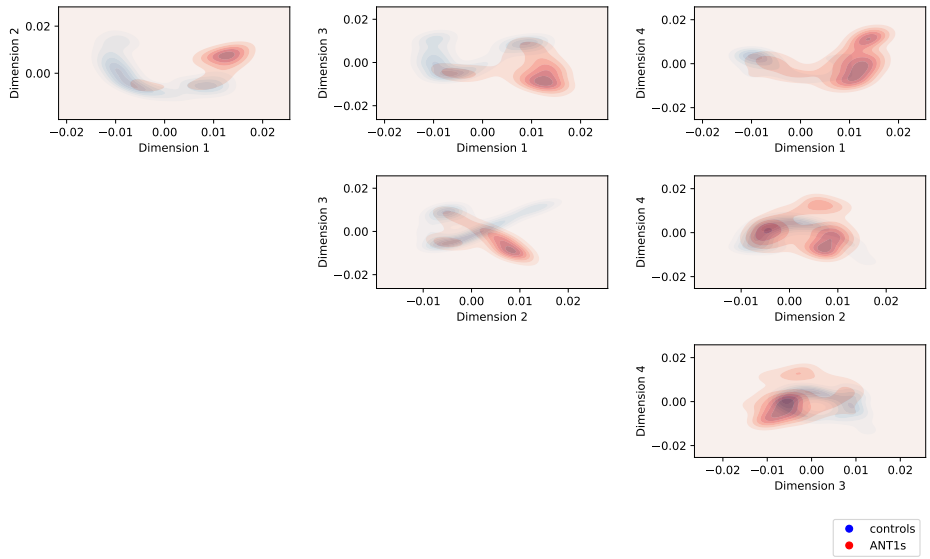


Figure 2.3: Probability density function for the control and ANT1 sample distributions, considering pairs of the first dimensions of the projected data.

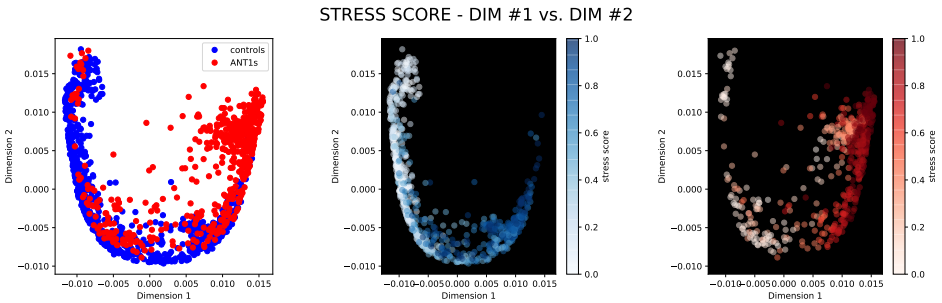


Figure 2.4: First two dimensions of projected data colored according to control (blue) and ANT1 (red) labeling (left). Separated distributions of control (middle) and ANT1 patient (right) samples, colormapped according to a stress score, consisting of the normalized HR sequence of each subject, mapped to the [0,1] interval.

Figure 2.3 displays the probability density functions learned from the distributions of control (blue) and ANT1 patient (red) samples, considering the pairwise combinations of dimensions 1-4, obtained using the non-parametric method of kernel density estimation [Epanechnikov, 1969]. We focused the analysis on the two first dimensions (column 1), as including further dimensions did not add any additional insight from a physiological perspective (see in 2.A an analogous analysis including dimension 3).

A scatter plot of the projected data, using the first two dimensions, is shown in Figure 2.4-left. In this plot, each point refers to a cardiac cycle of a subject, and is colored according to the control/ANT1 (blue/red) label. In the middle and right columns of the same Figure, we isolate each population's distribution of output-space samples and color them according to a stress score (computed as the HR value normalized by the minimal and maximal HR values of the corresponding patient). In both, the trend is to gradually transition from white (baseline/recovery) to dark blue/red (peak stress) in the counterclockwise direction. In fact, there is a continuum of response defined by the two distributions, where the ANT1 distribution is positively shifted in that same direction with respect to the control's.

To interpret the physiological implications of this spatial shift between distributions, we drew a path over the higher density regions and used MKR to reconstruct the velocity curves along such path. The path and points to backproject are shown in Figure 2.5-left, whereas the reconstructed mode of variation is shown in Figure 2.5-right.

Focusing on the evolution of the velocity curves from baseline to peak stress for the control population (see annotations in Figure 2.5-left), we observe a relative systolic lengthening, with the systolic peak happening later in the cycle, and a gradually shorter diastole, reaching some degree of E-A merging at peak stress. These are typical signatures of a normal response to exercise. Looking at the velocity patterns corresponding to the ANT1 baseline region (annotated in Figure 2.5), it is evident that ANT1 patients start off with some of these exercise signatures (e.g. shorter diastole); on the other hand,

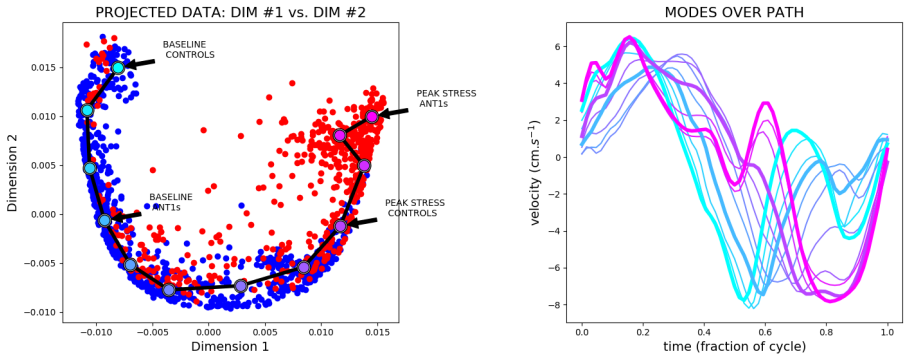


Figure 2.5: Distribution-based modes of variation. Left: Path over the distributions of control and ANT1 patients that were used for the estimation of the distribution-based modes of variation. The plotted points were the inputs for MKR. Right: MKR results, with color correspondence with the plotted path points in the left plot; the curves corresponding to the 4 annotated points are plotted with thicker linewidths.

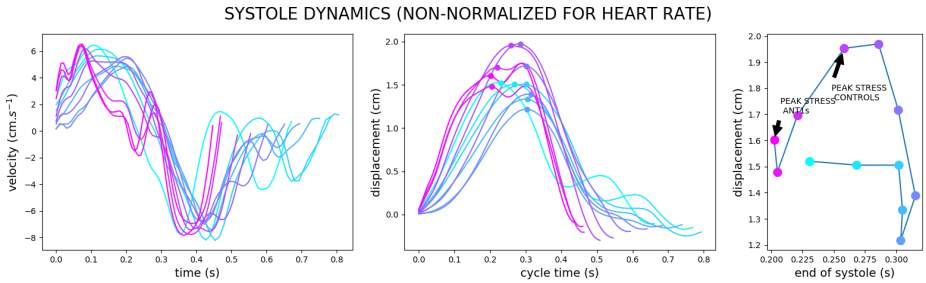
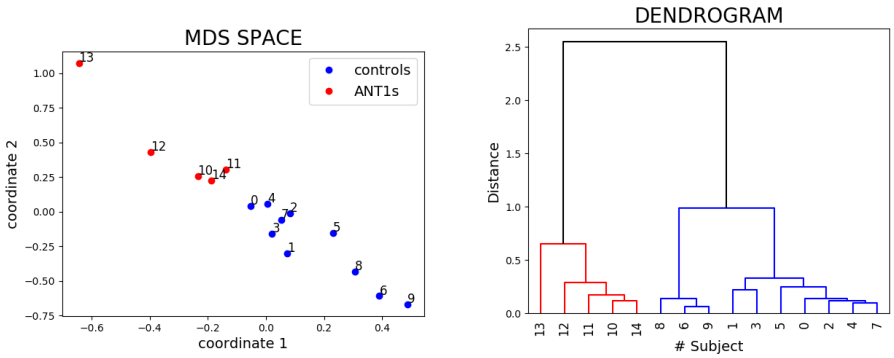


Figure 2.6: Systole dynamics. Left: velocity curves of Figure 2.5-right, non-normalized for HR. Middle: corresponding displacement curves. A marker is plotted on the systolic peak for each curve. Right: analysis of the timing and total displacement at the systolic peak, throughout the path drawn in Figure 2.5-left.

they also show a more accentuated augmentation of these signatures at peak stress (e.g. reaching complete E-A fusion), together with some additional shape changes in the velocity profile (especially noticeable



(a) (b)

Figure 2.7: Trajectory clustering based on DTW distance matrix (for these results, only the first dimension of the trajectories was considered). (a) 2D scatter plot of the subjects with MDS. (b) Hierarchical clustering (subjects 0-9 correspond to controls and subjects 10-14 correspond to ANT1 patients).

during systole). Focusing on systolic function in particular, we integrated the velocity curves to obtain the corresponding displacement profiles (Figure 2.6-middle), and plotted the timing of end-systole against the corresponding displacement (Figure 2.6-right). In these plots, the timing is not normalized for HR. It is observed that, as HR goes up, controls increase the peak contraction and, while the absolute ejection time reduces, the relative duration of systole with respect to the cycle length increases; on the contrary, ANT1 patients fail to modulate timings of events and contractility in the same manner, as they reach peak stress. The findings are in agreement with the results of [Sanchez-Martinez et al., 2017], that linked E-A fusion and reduction of contractility to exercise response in HFPEF, in a standard exercise context.

2.3.3. Sequence-wise analysis: subjects' trajectories

The trajectory defined by the projected samples of each particular subject in the output space (Figure 2.5-left) carries information regarding how that subject responds in terms of the patterns recovered in Figure 2.5-right. Thus, we took such trajectories as descriptors of response to stress, and used them to cluster the subjects into groups of response. For that, we computed the pairwise distances among all subject trajectories (using DTW, as described in 2.2.4.2), and used them as inputs to a hierarchical clustering algorithm. Other than this hard clustering, we used multidimensional scaling (MDS) to visualize a 2D scatter plot representation of the subjects, based on the same pairwise distances. Good clustering results were obtained even considering only the first dimension of the trajectories – in Figure 2.7a, the two clusters are even linearly separable, although there is a very subtle transition between them (MDS distributions accounting for more dimensions can be found in 2.A). In this scenario, the hierarchical clustering had perfect accuracy in the separation of healthy and diseased subjects (Figure 2.7b).

2.4. Discussion

We proposed an analysis framework for complex datasets composed of continuous multiview data sequences extracted from stress echo acquisitions, with the main objectives of *i)* discriminating healthy and pathological clusters of response and *ii)* understanding the underlying pathophysiological mechanisms. The framework extends the previous work by [Sanchez-Martinez et al., 2017] to nonstandardized stress echocardiography. Transitioning from standardized to nonstandardized data implies that each subject is no longer represented by a single point in the MKL output space, but by a variable number of points (with an associated time order). The main contribution of the proposed framework lies in the concept of studying low-dimensional

trajectories for clinical interpretation, an under-explored way to look at multiview clinical time series.

The discriminative power of the framework was first confirmed in Figure 2.3, where distinctive regions of higher control/ANT1 sample density were identified. It was again confirmed in Figures 2.7a and 2.7b, where a cluster analysis based on the subject trajectories in the output space accurately grouped them according to the diagnostic label.

Although there were distinctive regions of higher data density from the two populations, the two distributions were organized in a continuum of data. Such continuum was observed to be correlated with the stress level (Figure 2.4), and the ANT1 population data seemed to be positively shifted in the stress direction, when compared to the control distribution. In other words, the physiological patterns of the ANT1 population, at baseline conditions, are comparable to those found in controls during mild exercise.

To provide an idea of the variability found in trajectories (thus, signature intensities) within both control and ANT1 populations, we display the subject-wise projections in Figure 2.12 in 2.C.

While the studied populations used to illustrate the framework are distinctively different and easy to clinically discriminate based on heart failure symptoms at even the least exercise, our analysis can potentially provide novel insight in the physiology of this genetic mutation. However, the modest number of patients in the study precludes from any final conclusion when it comes to more in-depth pathophysiology analysis. On the other hand, the rarity of this mutation impedes the gathering of large datasets.

Among the main challenges of dealing with nonstandardized echocardiography sequences were those related to data processing and feature extraction, due to the complex nature of the data. Some recurrent problems that were especially likely to occur during stress were: noisy ECG, as in some cases the R peak became indiscernible and automated segmentation was not possible; out-of-plane heart motion resulting in an absence of Doppler signal, and significant breathing

motion relative to the defined region of interest. However, the final mapping to the low-dimensional space was found to be fairly robust to outliers (i.e. the obtained modes of variation/projections were not overfitting the outliers, e.g. cycles with saturation peaks, cycles with no velocity signal, badly segmented cycles due to bad quality ECG regions – which also gave origin to unphysiological values of HR, etc.), so there was no need to perform a preselection of cardiac cycles based on signal quality. This would probably not be the case if we did not have a fairly large ($\approx$ 1400 samples) dataset. The poor-quality cycles could be in most cases recognized based on the projection values in the output space (e.g. an overall poor-quality acquisition of subject 13 explains its considerable distance from the other subjects in the MDS plot (Figure 2.7a).

A high correlation between HR and the first dimension of the output space was found (0.81). To discard the possibility of HR strongly biasing the results, we repeated the whole experiments without feeding any HR information to the MKL algorithm (i.e. using only the velocity feature). After this, the correlation between HR and the first dimension of the output space remained high (0.67), with the main configuration of the two distributions and the corresponding modes of variation remaining similar (2.B). While the need of a multiview instead of a single-view dimensionality reduction algorithm could be arguable for this particular case, we still believe that taking HR as a feature to estimate the low-dimensional representation of the data can provide additional insight regarding the physiological interpretation and, in a scenario like ours, merging the two correlated features can add robustness when compared to pursuing a single-view approach on the velocity. Using more input features (e.g. velocity traces at other locations than the basal septum) would potentially allow a more specific characterization of the ANT1 response. Moreover, besides extending this analysis to velocity traces at other locations in the left and right heart chambers, one could also consider analyzing flow changes.

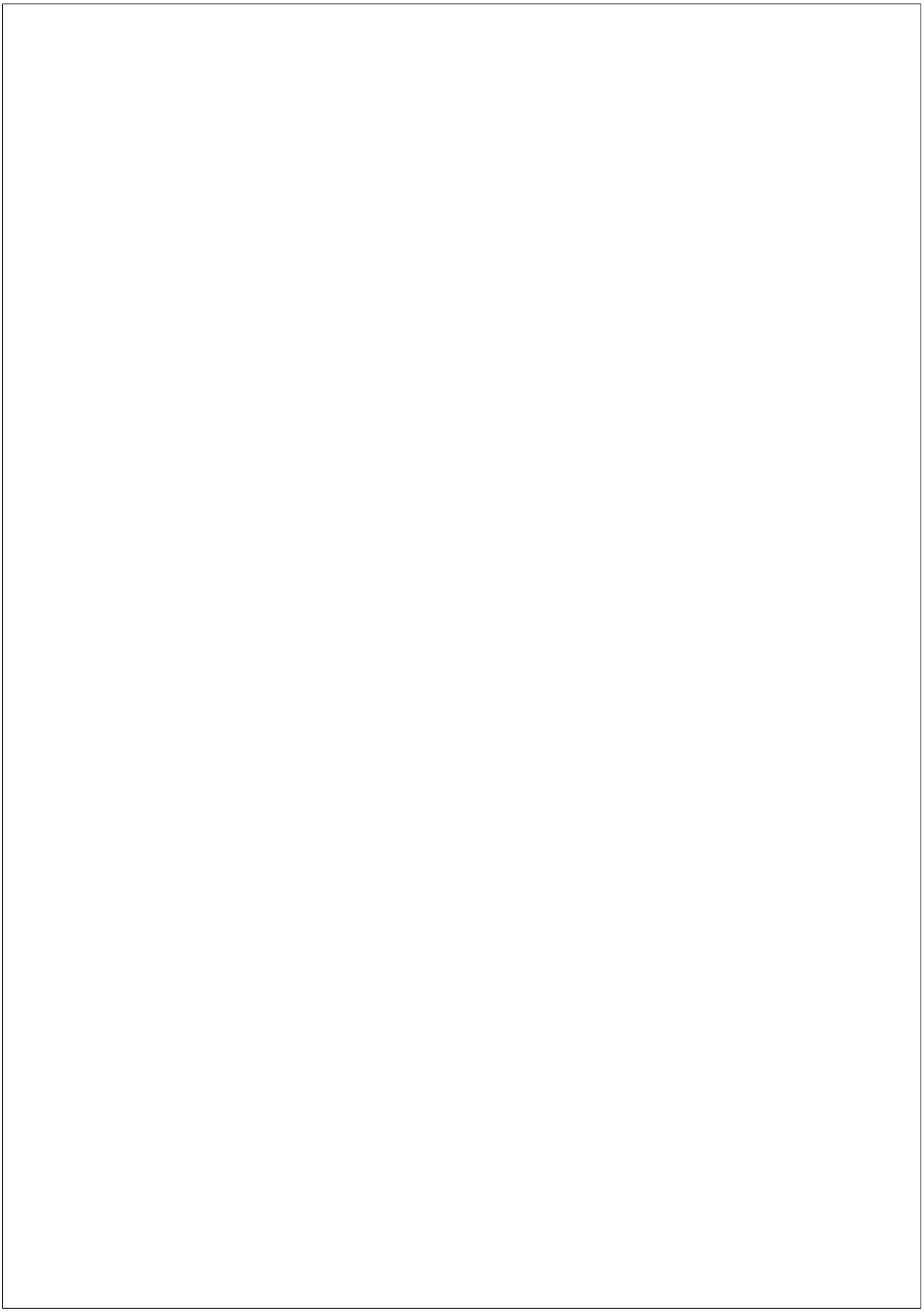
Despite the listed limitations, we demonstrated that the proposed

framework was able to reveal the progression of the pattern of response from the control to the pathological domain. The size of the dataset would not have permitted to come up with this pattern by visual inspection of the data.

The proposed framework can be flexibly adapted to the study of any given pathology, keeping in mind that the definition of the relevant set of features should be, naturally, carefully thought in a pathology-dependent manner.

2.5. Conclusion

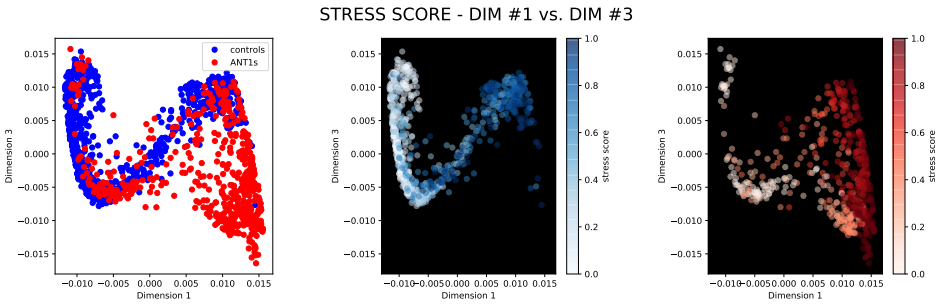
We have proposed a framework for the analysis of nonstandardized stress echocardiography sequence data. It uses unsupervised multiple kernel learning to merge myocardial velocity and heart rate information and obtain a low-dimensional representation of the data. The analysis is then performed in the new space, with multiscale kernel regression bridging the two spaces for interpretability. The framework is illustrated on handgrip exercise sequences acquired on a population of healthy controls and ANT1 mutation patients. The results show that the framework is able to detect distinctive clusters of response and provide insight into the underlying pathophysiological mechanisms, demonstrating its ability to handle this complex type of datasets, and the potential of nonstandardized protocols such as handgrip exercise for unmasking differential response mechanisms.



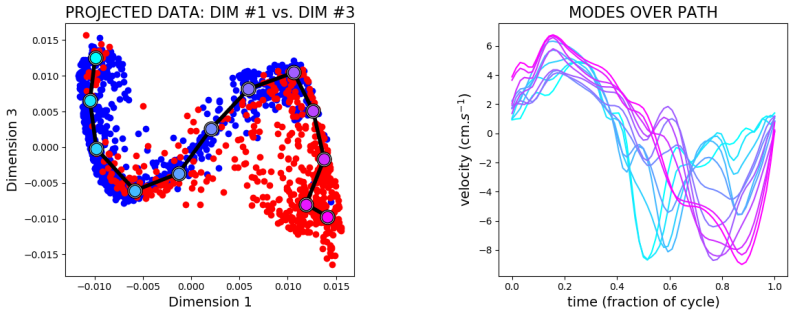
Appendices

2.A. Results for other combinations of output-space dimensions

Here, we show the sample-wise analysis when using dimensions 1 and 3 of the low-dimensional representation (Figure 2.8a), and 2 and 3 (Figure 2.9a). We also show the clustering of trajectories when using more than one dimension (Figure 2.10, middle and right).

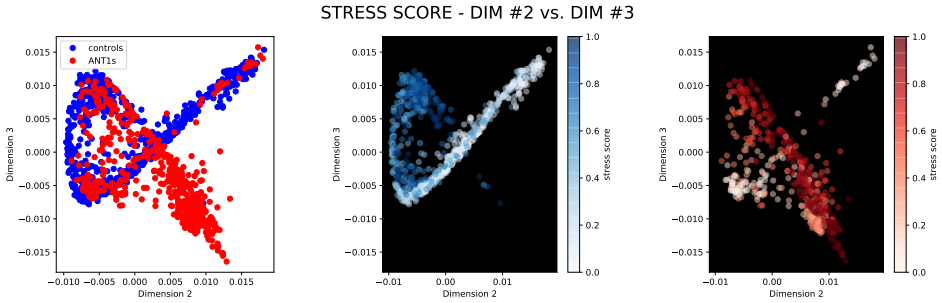


(a)

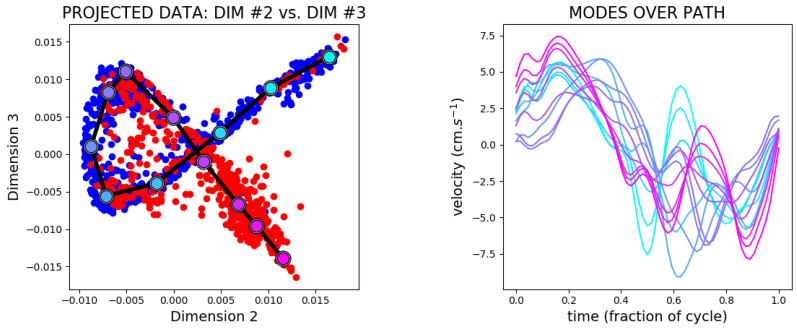


(b)

Figure 2.8: Analysis plots for the combination of output-space dimensions 1 and 3. (a) Regions of baseline, stress and recovery for the two populations; (b) velocity patterns reconstructed from the distribution.



(a)



(b)

Figure 2.9: Analysis plots for the combination of output-space dimensions 2 and 3. (a) Regions of baseline, stress and recovery for the two populations; (b) velocity patterns reconstructed from the distribution.

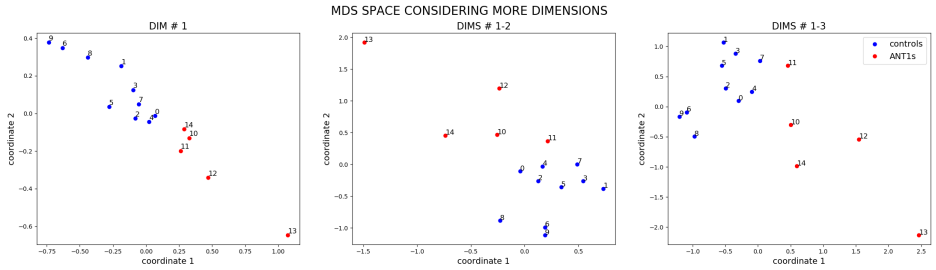


Figure 2.10: Changes in the MDS plots when considering more dimensions of the subject trajectories.

2.B. Experiments without HR

In this appendix we show the results of the sample-wise analysis using only velocity data (i.e. no HR data) as input to the MKL framework (Figure 2.11).

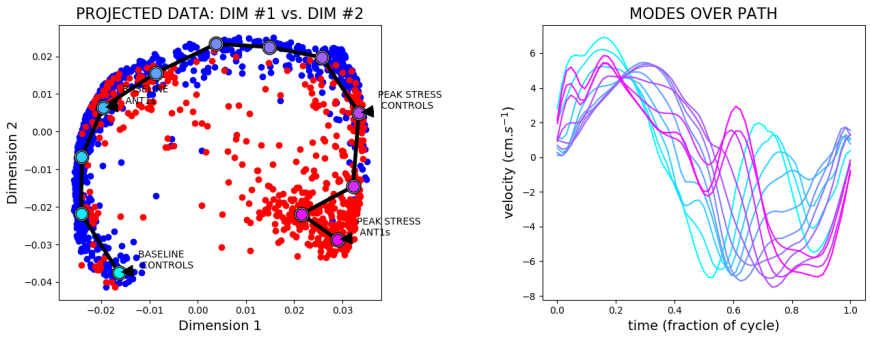


Figure 2.11: Experiments without considering HR as input feature of MKL: projected data and reconstructed mode of variation of the velocity feature.

2.C. Individual sequence lengths and projections

Herein, we detail individual sequence lengths and display the individual projections of the 15 subjects onto the 2d MKL space (Figure 2.12).

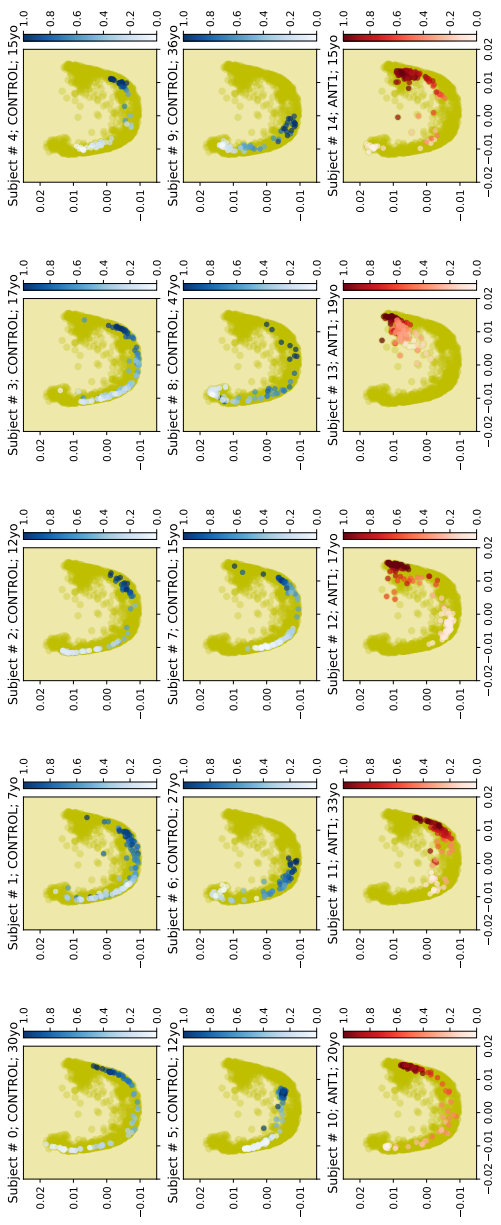
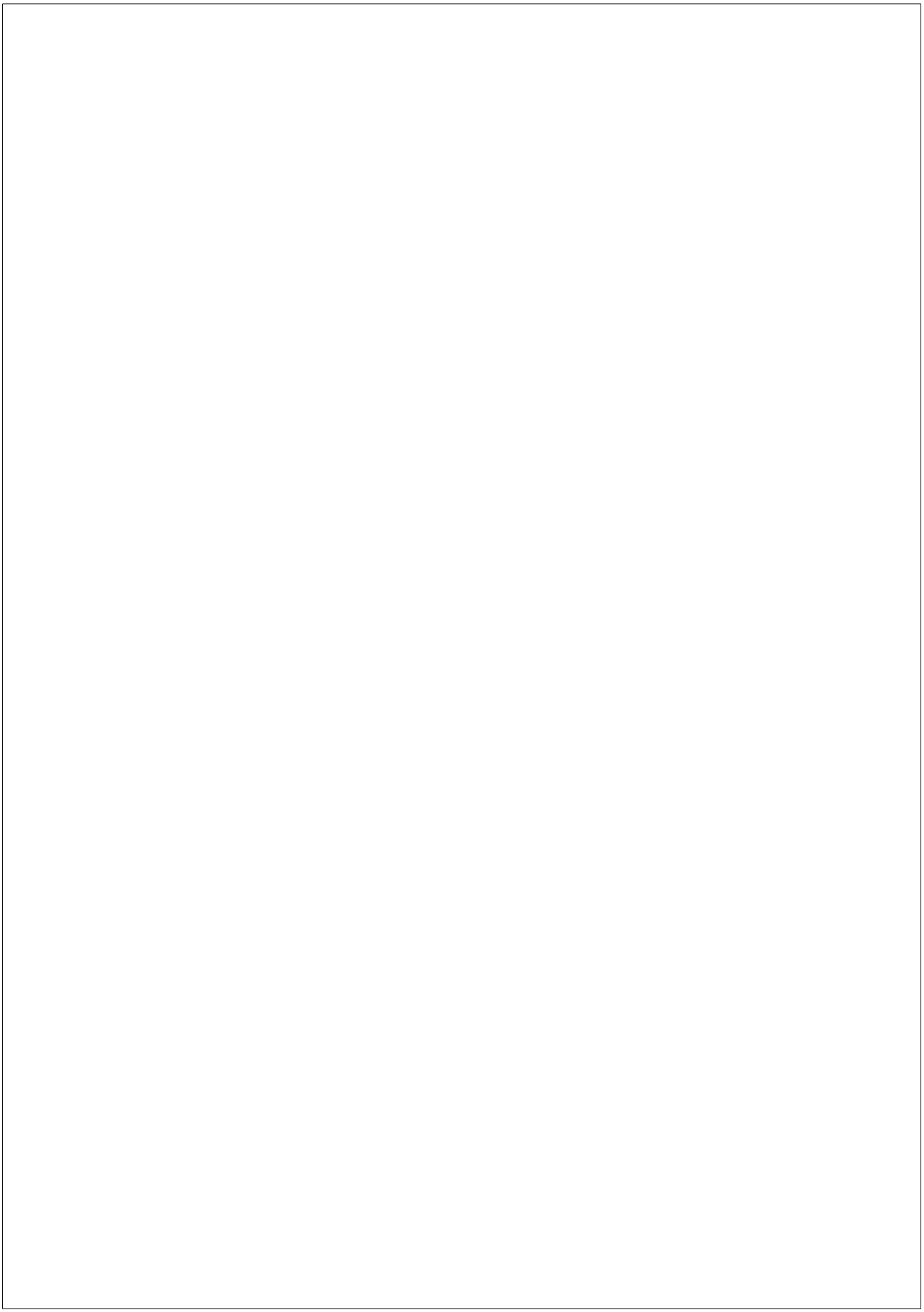


Figure 2.12: 2D MKL projection of the full dataset (dark green), superimposed by each subject's individual projection, color-mapped based on label (blue - controls; red - ANT1) and color-coded based on stress score, as previously done in Figure 2.4. Subject sequence lengths (from top left to bottom right): 74, 143, 64, 118, 64, 89, 63, 102, 67, 58, 95, 88, 110, 137, 105.



Chapter 3

A PERSONALISED APPROACH FOR EFFECTIVE LABOUR MONITORING BASED ON MACHINE LEARNING ASSESSING WOMEN'S SIMILARITY AND OPTIMAL TEMPORAL PROGRESSION

This chapter is adapted from: M. Nogueira, G. Piella, M. De Craene, C. Yagüe, S. Sanchez-Martinez, P. Martí, M. Bonet, O.T. Oladapo, B. Bijnens. A personalised approach for effective labour monitoring based on machine learning assessing women's similarity and optimal temporal progression. *Submitted to Nature Methods.*

3.1. Introduction

Infant as well as maternal death rates are still unacceptably high, especially in low-income countries [World Health Organization, 2015, Oladapo et al., 2015, Souza et al., 2015, Yang et al., 2017]. Although early fetal or late neonatal mortality occur, most deaths and severe morbidities have origin around the time of childbirth, making quality of care during this period critical for a positive outcome [Oladapo et al., 2015]. However, there is still little consensus on what are the best approaches to labour monitoring and decision making, and actual practice is very diverse [Robson et al., 2015]. Within this diversity, some worrisome patterns have been identified. On the one hand, there is a seemingly unjustified global escalation of rates of labour intervention, especially caesarean sections (CSs) [World Health Organization,]. CSs bear risks for the mother, baby, and future pregnancies, which are intensified in women with low access to adequate care [Betrán et al., 2015]. On the other hand, in low-resource environments, interventions can be limited to suboptimal rates. In general, there are inequalities in intervention rates among and within countries – which correlate with economic inequalities – with some women receiving too little intervention when needed, and others receiving too much [Boatin et al., 2018].

An ideal standardized practice would always have as first objective the minimization of adverse outcome, while also avoiding unnecessary interventions, as a way to minimize risks and optimize the management of resources.

The closest to a reference labour monitoring and decision support tool has been World Health Organization (WHO)'s partograph, where the healthcare providers plot the evolution of multiple maternal and fetal measurements over the course of labour, against predefined ranges of "normality". A central feature of the partograph is the cervicograph, which plots cervical dilatation over time against recommended "alert" and "action" lines. The latter were designed to guide healthcare providers in decision making when cervical dilatation progresses at

a slower pace than a reference value (1 cm/h), after the onset of 4 cm (up until recently interpreted as the onset of the active phase of labour). However, almost three decades after the issuance of the partograph, there is (1) increasing scepticism regarding the use of alert and action lines as reference for all labours, regardless of women characteristics (e.g. age, ethnicity, socioeconomic context), obstetric history (e.g. previous pregnancies, previous CS) and other factors that could redefine "normality" in spontaneous labour progression, (2) lacking evidence of positive impact of its use on outcome, and (3) disappointingly low rates of appropriate use [Oladapo et al., 2015].

Recent studies have put effort in evaluating the predictive value of different partograph variables regarding severe maternal and fetal adverse outcomes [Bonet et al., 2019, Robson et al., 2015, Souza et al., 2018, Oladapo et al., 2018], essentially by assessing how the crossing of predefined thresholds, including the alert and action lines of the cervicograph, correlated with those outcomes. In some cases, customized versions of the alert and action lines were used for different obstetric groups, such as based on the Robson's 10-group classification [Robson et al., 2015, Souza et al., 2018, Oladapo et al., 2018]. The reported predictive performance was considered poor in all cases, suggesting that a univariate, transversal approach to the partograph is not effective in the prediction of severe adverse outcome.

As more complete datasets and more sophisticated data analysis tools become available, developing data-driven monitoring and decision support systems (DSSs) becomes an appealing option. The WHO itself expressed an interest in exploring evidence-based alternatives with the Simplified, Effective, Labour Monitoring-to-Action (SELMA) [Souza et al., 2015] project.

Most of the proposed DSSs in clinical medicine have been developed using data collected in a well controlled setting, with a standardised data acquisition protocol, and often specific interventions with clear indications. However, in a realistic clinical setting, given the complex nature of labour monitoring with non-standardised intervention decision making and the need to detect crucial problems

that only rarely occur, the task becomes quite challenging. Some of the biggest challenges for a DSS are: (1) the wide variety in initial presentation of the pregnant women in the hospital, followed by the temporally varying progress, makes the direct application of typical learning techniques challenging, and the fact that temporal samples are recorded at nonstandardized timings complicates any analysis further; (2) intervention and outcome data reflect a site- and study-specific practice, which does not necessarily align with the one that minimizes adverse outcome and unnecessary intervention; (3) very low rates of adverse clinical outcomes generate severely imbalanced data, which complicates predictive learning.

Few data-driven frameworks for labour monitoring and decision support have yet been proposed. Most research effort went into the specific subproblem of prediction of CS, towards decision support [Souza et al., 2019, Burke et al., 2017, Chen et al., 2004, Campillo-Artero et al., 2018, Levine et al., 2018, Janssen et al., 2017, Harper et al., 2013]. In most cases, multivariate prediction models are built using logistic regression (supervised learning). The majority of models use admission-time data only [Souza et al., 2019], providing an initial estimate of risk of CS that is not dynamically updated.

Souza et al. [Souza et al., 2019] compared the predictive performance of (1) admission-time models, (2) "interval" models – the 6 hours after the onset of 4 cm of cervical dilatation were divided in 2-hour intervals, and a different model was learned with updated intrapartum measurements of each interval – and (3) maximum score models – trained with the "maximum scores" of the dynamic descriptors (extreme values achieved throughout the whole course of labour or, in some cases, final values). The WHO's SELMA dataset [Souza et al., 2015] was used in all cases. The maximum score models outperformed all others; however, they were trained and tested on information transversal to the whole labour duration, and how they would perform in a real-time scenario is unknown. The admission-time models were the least performing, and the interval models increased their performance from first to last interval, in between the admission-

time and maximum-score model performances. This is a somewhat intuitive result: updates in intrapartum data are relevant for the decision. However, the fact that total sample size decreases from first to last interval, while CS is more likely in later stages, should not be overlooked when interpreting the results. Of all models, only the interval models could support real-time decision-making. A foreseen challenge in the integration of these models in a decision support tool is the management of the interval prediction by the healthcare provider. At each interval, the models provide a binary prediction (or a probability) of CS; thus, if the model predicts a CS, how to temporally handle the true positive is not obvious, since each interval model learned if CSs occurred, regardless of the interval in which they actually occurred, not being able to relate their prediction with a time of decision/incision. On the other hand, the significant differences in the models' performances at each interval make it even more challenging (for the healthcare provider) to manage the probability/prediction information.

To the best of our knowledge, the real-time monitoring component has not been the priority in recent research. Indeed, little effort has been made towards simplifying the visualization and interpretation of what can be an overwhelming amount of dynamic information required for healthcare providers' decision-making, along with any evidence-based suggestions for prognosis and interventions.

The objectives of this paper are (1) to present a novel framework for labour monitoring based on interpretable machine learning and (2) to showcase its potential as a basis for a decision support system.

With regard to objective (1), we aim to provide a personalized dynamic labour monitoring-to-action-tool that (i) accounts for all variables and their interactions, (ii) defines "normality" in labour progression in a personalized manner, and (iii) also provides personalized prognosis of intervention and outcome, and their most likely timing, based on study-specific evidence. This way, we address the limitations of the current univariate, generic approach to the implementation of the partograph, and enrich its value with study-based knowledge,

while maintaining clinical familiarity and interpretability.

With regard to objective (2), a preliminary assessment of the framework's potential as basis for a DSS is performed. For validation, we compare performance with the current state of the art – the partograph and the prediction models of Souza et al. [Souza et al., 2019]. In all cases, the SELMA dataset is used for illustration and performance comparison. Lastly, under the premise that practice can be rather heterogeneous, we additionally perform "subgroup" analyses, in order to assess whether prediction is an easier task for some subgroups of the population than others. The first subgroup analysis addresses a specific result of recent observational studies linking admissions in the active phase of labour (using 4 cm as onset cut-off) with lower likelihoods of labour interventions, without increasing maternal or perinatal morbidity – suggesting that early admission *per se* increases the chance of intervention [Holmes et al., 2001, Neal et al., 2014, Bailit et al., 2005, Mikolajczyk et al., 2016, Chuma et al., 2014]. In an attempt to reduce the excess of intervention in early-admission groups, the WHO has recently updated the definitions of latent and active phases of labour, now recommending 5 cm as a more appropriate cut-off. However, this happened after the SELMA study was conducted. A second subgroup analysis explores practice differences in a more unsupervised way by dividing the population in several clusters of similar women based on admission characteristics.

3.2. Methods

3.2.1. Data and Preprocessing

The WHO's SELMA dataset was used to illustrate the proposed framework. It comprises information from 9995 deliveries across 13 different facilities across Nigeria and Uganda. A very complete set of features including demographics, medical history and previous pregnancy information are collected at admission, followed by the first (baseline) maternal and fetal assessments that are to be monitored

during the course of labour. Intrapartum updates of these values are available at non-standardized time intervals. A detailed set of intra- and post-partum complications, interventions and outcomes is also available. We refer to the features that remain unchanged during labour (e.g. demographics and medical history) as *static* features, and those to be monitored continuously as *dynamic* features.

We used 52 features (33 static and 19 dynamic) to characterize women in labour at any time point. Some of them were directly taken from the original SELMA dataset, whereas others result from the combination of several of the original features. The features are described in Tables 3.5 and 3.6 of Appendix 3.A, with those which were in some way engineered having their names emphasized in bold.

Often, some values were missing from different features/follow-ups of each woman. At admission, missing values in features regarding history of lung disease, emotional and painful distress were interpreted as absence of abnormality, and missing data on axillary temperature and number and duration of contractions were imputed with the value of the first follow-up (after admission). In the case of temperature, if the first-follow-up value was also missing, the average temperature at admission was assumed. After these operations, 549 women still presented important missing admission data and were discarded from the analysis. Another 876 women were removed from the analysis due to time inconsistencies. For the remaining 8470 women, missing data among follow-ups was dealt with through previous (follow-up) value propagation. The way the SELMA dataset is used to illustrate and evaluate the proposed framework is detailed in Section 3.2.3.

3.2.2. Framework

Figure 3.1 illustrates the proposed approach, from evidence data to DSS. We first use manifold learning to represent multivariate study data in a lower-dimensional, interpretable space, where subjects are positioned based on their similarities, and temporal data are visualized as low-dimensional trajectories. New subject infor-

mation is handled by (1) projecting updated subject data to this space, (2) retrieving "similar" study subjects, i.e., those confined to a close neighborhood, (3) taking those who underwent spontaneous, complication-free labours to estimate a normal/expected progression and to calculate the subject's deviation from it, and (4) taking the ratios of different interventions/outcomes among all retrieved neighbours as estimates of chance of occurrence. We thus "personalize" monitoring and decision support by redefining "normality" and risk estimates based on the labour progress, intervention and outcome data of "peers" (i.e. similar subjects within the study population).

The core objective of this paper is to present our framework as this high-level pipeline, leaving room for flexibility in implementation. In this section, we describe a possible implementation, which we use for the purposes of illustration and evaluation with the SELMA dataset.

To obtain the similarity-ruled space, we use unsupervised multiple kernel learning (MKL) [Lin YY, 2011, Sanchez-Martinez et al., 2017, Sanchez-Martinez et al., 2018, Nogueira et al., 2020a], an algorithm that allows representing heterogeneous features in a unified manner and subsequently merging their information to learn a lower-dimensional embedding of the data where samples are spatially ordered by similarity.

3.2.2.1. Learning the Projection Model and Precomputing Projections Database

A necessary first step of the framework is learning a projection model to a similarity-ruled space from the study data and using the newfound projection model to precompute a database of projections (Figure 3.1, top). In this paper, this is achieved using an unsupervised MKL algorithm.

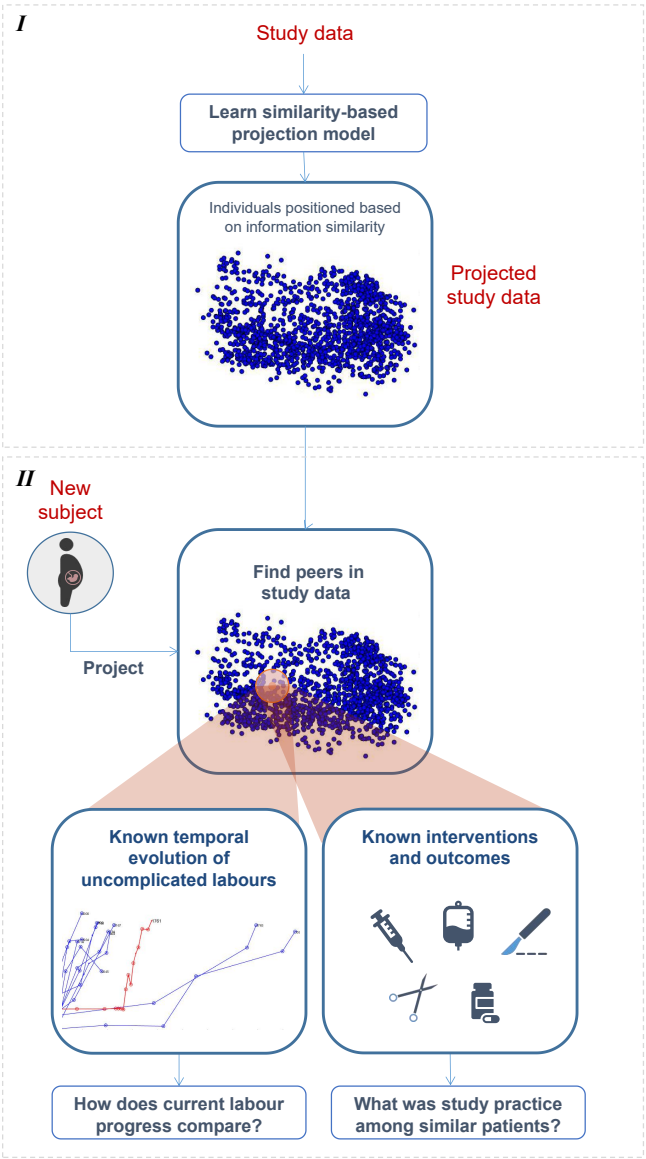


Figure 3.1: High-level illustration of the proposed framework. Stage I - learning a similarity-based projection model and precomputing a database of projections from study data. Stage II - peer-based monitoring and decision support.

Overview of MKL

Let us consider N data samples, each described by M uni- or multidimensional features. An MKL projection to a D-dimensional space is parameterized by a projection matrix $A \in \mathbb{R}^{N \times D}$ and a vector $\beta \in \mathbb{R}^M$ that contains the weight of each feature in the mapping. Instead of operating directly on the raw data, A and β operate on kernelized (similarity) data. Let x_i^m denote the data associated with the m^{th} feature of the i^{th} data sample, with $i = 1, ..., N$ and $m = 1, ..., M$. In this paper, all features are unidimensional, so $x_i^m \in \mathbb{R}$. Additionally, let us use the simplified notations $x^m \in \mathbb{R}^N$ for the vector of values of feature m for all N samples, $x^m = (x_1^m, ..., x_N^m)^T$, and $x_i \in \mathbb{R}^M$ for the vector of M feature values of sample i, $x_i = (x_i^1, ..., x_i^M)^T$. Different data types may be associated with different notions of similarity. In this paper, we adopt the kernel functions proposed in [Daemen and De Moor, 2009] for clinical data. Let k^m denote the kernel function associated with feature m. For continuous/ordinal variables, the similarity between input samples i and j is measured by

$$k^m(x_i^m, x_j^m) = 1 - \frac{|x_i^m - x_j^m|}{\max x^m - \min x^m} \quad , \tag{3.1}$$

whereas for nominal variables,

$$k^m(x_i^m, x_j^m) = \delta(x_i^m - x_j^m) \quad , \tag{3.2}$$

with δ the Kronecker delta function.

Let $K \in \mathbb{R}^{N \times N \times M}$ denote the three-dimensional matrix whose entries are $K_{ijm} = k^m(x_i^m, x_j^m)$. with $i, j = 1, ..., N$ and $m = 1, ..., M$. Let $K_i \in \mathbb{R}^{N \times M}$ denote the i^{th} slice of K along the first dimension,

$$K_i = \begin{pmatrix} k^1(x_1^1, x_i^1) & ... & k^M(x_1^M, x_i^M) \\ ... & ... & ... \\ k^1(x_N^1, x_i^1) & ... & k^M(x_N^M, x_i^M) \end{pmatrix}, \tag{3.3}$$

and $K^m \in \mathbb{R}^{N \times N}$ the m^{th} slice along the third dimension

$$K^m = \begin{pmatrix} k^m(x_1^m, x_1^m) & ... & k^m(x_1^m, x_N^m) \\ ... & ... & ... \\ k^m(x_N^m, x_1^m) & ... & k^m(x_N^m, x_N^m) \end{pmatrix}. \tag{3.4}$$

In short, K_i encodes the similarity coefficients among sample i and all other samples (rows), in terms of all M features (columns). On the other hand, K^m is a symmetric matrix that encodes the pairwise similarities of all N samples according to feature m.

In our unsupervised MKL model, the projection $y_i \in \mathbb{R}^D$ of $x_i \in R^M$, with $D \leq M$, becomes a function of K_i:

$$y_i = A^T K_i \beta \quad . \tag{3.5}$$

The MKL problem is then formulated as

$$\min_y \sum_{i,j} \|y_i - y_j\|^2 W_{ij} \quad , \tag{3.6}$$

$$\text{s.t.} \sum_i \|y_i\|^2 W'_{ii} = 1 \quad , \tag{3.7}$$

where W is an affinity matrix, computed as a (linear or non-linear) combination of all $\{K^m\}_{m=1}^M$. Thus, each entry W_{ij} encodes the similarity between samples i and j based on contributions from all features. In this paper, W is computed as the average of all K^m. The minimization imposes that samples that are similar in the input space (high W_{ij}) are mapped to close positions in the output space. Constraint (3.7) removes an arbitrary scaling factor in the output embedding and eliminates trivial solutions, with $W'_{ii} = \sum_j W_{ij}$. Plugging (3.5) into (3.6) and (3.7) the problem translates into finding A and β such that

$$\min_{A,\beta} \sum_{i,j} \|A^T K_i \beta - A^T K_j \beta\|^2 W_{ij} \quad , \tag{3.8}$$

$$\text{s.t.} \sum_i \|A^T K_i \beta\|^2 W'_{ii} = 1 \quad . \tag{3.9}$$

In practice, A and β are found by iteratively solving a generalized eigenvalue problem for A, and a semidefinite programming problem for β [Lin YY, 2011]. Once the model parameters (A, β) are estimated, the projection $y_u \in \mathbb{R}^D$ of a new sample $x_u \in \mathbb{R}^M$ amounts to a generalization of eq. (3.5):

$$y_u = A^T K_u \beta \quad , \tag{3.10}$$

where $K_u \in \mathbb{R}^{N \times M}$ and $K_{ium} = k^m(x_i^m, x_u^m)$, with $i = 1, ..., N$ and $m = 1, ..., M$.

Application to our problem

We use the unsupervised MKL to learn the projection model of the study dataset. Let us consider our database consists of admission and follow up data of P subjects. That is, each subject p (with $p = 1, ..., P$) has a sequence of time points $f = 0, ..., F_p$, F_p being the number of follow-ups of subject p and $f = 0$ corresponding to the first assessment (admission). Let t_f^p denote the timing of follow-up f of subject p, computed as the absolute time in hours since admission ($t_0^p = 0$ hours), x_{p,t_f^p} the corresponding data sample, and y_{p,t_f^p} the corresponding MKL projection.

3.2.2.2. Monitoring a New Subject

Here, we describe how the model and projections computed in Section 3.2.2.1 are used in the dynamic monitoring of a new subject (Figure 3.1, bottom). Given a new subject q at follow-up f:

1. Update subject. That is, project the data sample $x_{q,t_f^q} \in \mathbb{R}^M$ to $y_{q,t_f^q} \in \mathbb{R}^D$ using eq. (3.10), i.e., $y_{q,t_f^q} = A^T K_{q,t_f^q} \beta$.

2. Find peers. Peers are defined as the study subjects whose projections at time t_f^q are within a limited neighborhood of y_{q,t_f^q}. In practice,

- For each subject $p = 1, ..., P$, retrieve the projection at time t_f^q, y_{p,t_f^q}. Since many study subjects will not have follow-up data specifically at t_f^q, the value of y_{p,t_f^q} is obtained for each subject by linear interpolation on the precomputed $y_{p,t_f^p}, f = 0, ..., F_p$.

- Study subject p is considered a peer of q at time t_f^q if

$$\sum_{d=1}^{L} \left(y_{p,t_f^q}^d - y_{q,t_f^q}^d \right)^2 \leq R^2, \quad L \leq D, \tag{3.11}$$

  i.e., if the projection of subject p is contained within a hypersphere of dimensionality L and radius R centered on that of subject q. To account for scaling differences among dimensions d, condition (3.11) is computed on a standardized form (zero mean and unit standard deviation for all dimensions) of the projection data.

3. Estimate deviation from ideal progression Let $\mathbb{H}$ denote the set of peers of subject q at time t_f^q, i.e. the subset of study subjects obeying (3.11). Additionally, let $\mathbb{S}$ represent the subset of the P subjects whose labour progressed spontaneously towards ideal outcome, i.e. without any complications or interventions.

- The t_f^q update of the estimate of normal progress for time $t > t_f^q$, $E_{t_f^q}(t)$, is given by

$$E_{t_f^q}(t) = \frac{1}{|\mathbb{C}(t)|} \sum_{p \in \mathbb{C}(t)} y_{p,t}, \quad \mathbb{C}(t) = \{p \mid (p \in \mathbb{H} \cap \mathbb{S}) \wedge (t_{F_p}^p \geq t)\} \tag{3.12}$$

  corresponding to the mean of the projections of all peers with normal labour at time t (provided they exist or can be interpolated for such timing). The corresponding standard deviation $\sigma_{t_f^q}(t)$ is computed as an estimate of "normal" variability.

- In the next follow-up, at time t^q_{f+1}, we can verify how much the projection $y_{q,t^q_{f+1}}$ deviates from the predicted "normal" position, $E_{t^q_f}(t^q_{f+1})$. Specifically, for each dimension d, we quantify deviation from normality as the z-score

$$z^d_{t^q_{f+1}} = \frac{y^d_{q,t^q_{f+1}} - E^d_{t^q_f}(t^q_{f+1})}{\sigma^d_{t^q_f}(t^q_{f+1})} \quad . \tag{3.13}$$

4. Predict (timings) of interventions/outcomes Let us refer to interventions and outcomes as events. Let e denote an event of interest, $\mathbb{E}$ the set of study subjects that experienced that event and $t^{p,e}$ the timing of such event for subject $p \in \mathbb{E}$.

- The chance of e, at time t^q_f, is updated as

$$\pi^e_{t^q_f} = \frac{|\mathbb{G}|}{|\mathbb{H}|}, \quad \mathbb{G} = \{p \mid (p \in \mathbb{H} \cap \mathbb{E}) \wedge (t^{p,e} \geq t^q_f)\} \tag{3.14}$$

  i.e., the ratio of peers that experienced e at $t \geq t^q_f$.

- Moreover, a probability density function can be fitted to the distribution of all $t^{p,e}, p \in \mathbb{G}$, so as to obtain an estimate of the probability of e with respect to time, $\pi^e_{t^q_f}(t), t \geq t^q_f$.

3.2.3. Performance evaluation

We evaluate whether the proposed framework is capable of capturing relevant information in terms of predictive value regarding the occurrence of events of interest. To this end, we apply our framework to the SELMA study and assess the predictive performances of simple and intuitive descriptors.

In this paper, the evaluation is focused on the prediction of CS and severe adverse (bad) outcome (BO), the most represented challenges in the literature. However, the framework can be used for any event

of interest. Uncomplicated labour was defined as those with no occurrences of amniotomy, labour augmentation, CS or BO. BO was defined consistently with BO definitions in previous literature regarding the SELMA study [Souza et al., 2018, Oladapo et al., 2018, Bonet et al., 2019], as the composite of: stillbirth, intra-hospital early neonatal death, neonatal use of anticonvulsants, neonatal cardiopulmonary resuscitation, Apgar score below 6 at 5 minutes, uterine rupture, maternal death or organ dysfunction preceded by dystocia.

Given a subject q, three predictors were defined per targeted event $e \in \{CS, BO\}$:

1. Chance estimate (as compared to the occurrence in the SELMA study) :

$$v_{\pi}^{e} = \max_{f} \pi_{t_f^q}^{e}, \quad f = 1, ..., F_q. \tag{3.15}$$

2. Combination of chance estimate with deviation from normality as calculated by our framework:

$$v_{\pi z}^{e} = \max_{f} \left[\pi_{t_f^q}^{e} * \max_{d} |z_{t_f^q}^{d}| \right], \quad f = 1, ..., F_q, \quad d = 1, ..., D. \tag{3.16}$$

3. Combination of chance estimate with deviation from normality and time since admission:

$$v_{\pi z t}^{e} = \max_{f} \left[\pi_{t_f^q}^{e} * \max_{d} |z_{t_f^q}^{d}| * t_f^q \right], \quad f = 1, ..., F_q, \quad d = 1, ..., D. \tag{3.17}$$

After the preprocessing steps described in 3.2.1, the remaining study subjects ($n = 8470$) were randomly assigned to a training ($n = 6349$) and testing set ($n = 2121$) (step 1 in Figure 3.2). The rates of occurrence of events of interest were verified to be balanced

between the two sets and representative of the whole population's ($\approx 13\%$ for CS and $\approx 2\%$ for BO). The admission-time data of the training set were used in the learning of the MKL projection model (step 2), and the model was used to project all training and testing data (step 3).

The projected training set was further divided in three folds, and the framework was run three times in cross-validation style, i.e. each time with two of the folds simulating the projection database and the remaining fold simulating "new subjects" (step 4). The predictors $\{v_k^{CS}\}$ and $\{v_k^{BO}\}, k \in \{\pi, \pi z, \pi zt\}$, were then collected for all the "new subjects" of each experiment. For each descriptor-label combination, a receiver operating characteristic curve was plotted and its area (AUC) computed. The predictor cut-off values that maximized the joint entropy of the sensitivity (SE) and specificity (SP) were selected, so as to favor balanced solutions. Other performance metrics such as the positive and negative predictive values (PPV/NPV) were also computed for the selected cut-off values. A different cut-off might be chosen in a clinical DSS depending on the targeted focus on reducing false negative or positive prediction.

The framework was then run with the full training set simulating the projections database and the testing set simulating "new subjects" (step 6). The threshold values learned in the cross-validation stage were applied to the set of descriptors collected for the women in the testing set, to infer about generalizability (step 7).

As previously referred, for validation, performance was compared to those of the partograph's alert and action lines. The alert and action lines are cut-offs by definition, and their predictions are based on them being crossed or not. Specifically, if cervical dilatation between 4 cm and 10 cm happens at a slower pace than 1 cm/h, the alert line is crossed. The action line is parallel to the alert line, only shifted 4 h to the right [Souza et al., 2018, Oladapo et al., 2018]. In the case of the CS prediction problem, performance was also compared with those of the prediction models by Souza et al. [Souza et al., 2019].

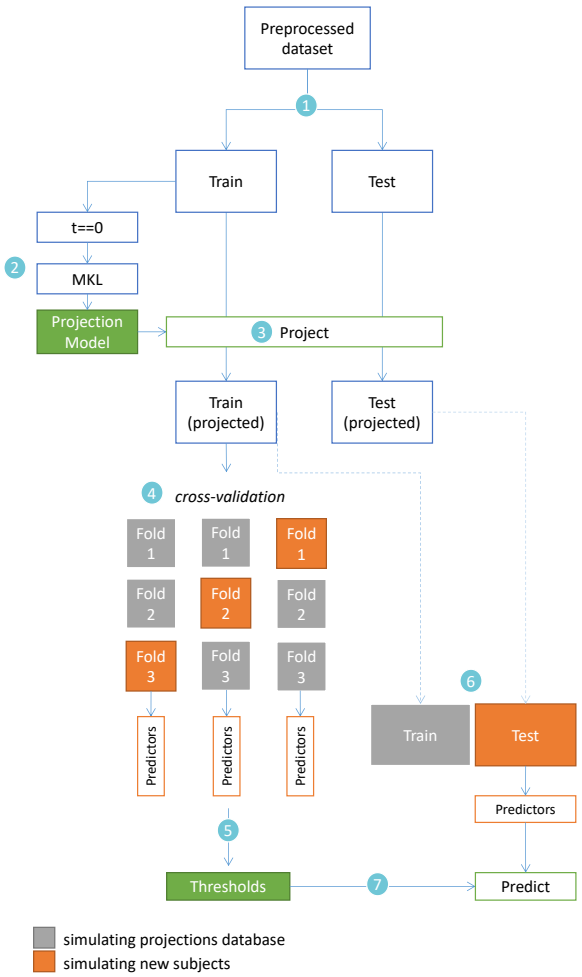


Figure 3.2: Evaluation of the proposed framework using the SELMA dataset. 1 – train/test partition; 2 – learning the MKL projection model with the admission-time features of the training set; 3 – projecting all training and testing data; 4 – three-fold cross-validation and predictor extraction with the training set; 5 – extraction of cut-off values for the predictors; 6 – framework application and predictor extraction with the testing set; 7 – application of learned cut-offs in the testing set predictors.

Lastly, the subgroup analyses were performed. For the first, the population was split into early- and late-admission subgroups. For the second, the population was divided in several subgroups of women with similar admission characteristics.

3.3. Results

Results hereby shown correspond to experiments ran for $M = 52$ features, $D = 10$ dimensions, $L = 4$ dimensions and $R = 0.5\mu_0$, where μ_0 denotes the average pairwise subject projection distance in the database at $t = 0$ (a grid search was performed over D, L and R to tune values based on performance). These values alone, however, are not very informative regarding the actual number of peers that each test subject is being compared to. To gain some insight on this matter, let us consider the final experiment (step 6 of Figure 3.2), where the framework is ran for 2121 test subjects, with 6349 subjects in the projections database. We plot the histogram of the numbers of peers of all test subjects (Figure 3.3) at $t = 0$ (left) and considering all follow-ups (right). At $t = 0$, a test subject is compared, on average, to 573 subjects, corresponding to 9% of all available training subjects. When all follow-ups of all test subjects are accounted for, the average value drops to 370 subjects. This is an expected effect given that the number of available training subjects decreases as time advances.

3.3.1. The similarity-ruled space and clinical interpretability

Figure 3.4 illustrates the similarity-based spatial ordering in the MKL space obtained for the SELMA dataset and the clinical interpretability of the obtained projections. The plots depict the projections of the samples used to learn the MKL model, i.e. the admission-time data samples of the training set subjects. Each scatter point thus corresponds to one subject, and the whole scatter plot is a snapshot of

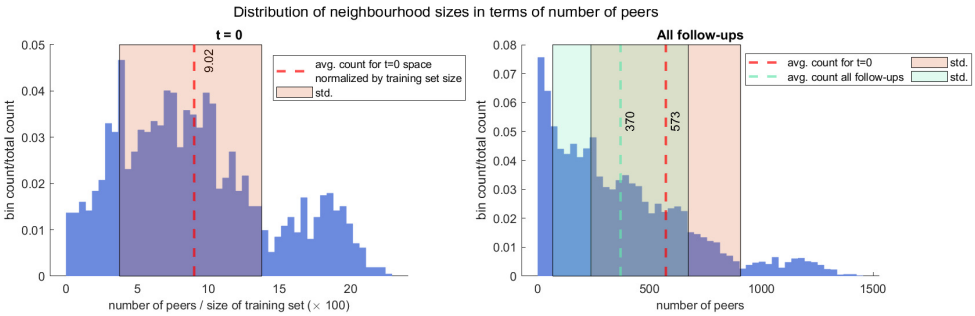


Figure 3.3: Distribution of neighbourhood sizes for step 6 of Figure 3.2. Left: at $t = 0$, in terms of percentage of the total number of subjects of the training set. Right: considering all follow-ups, in absolute number.

the projections of the training subjects at $t = 0$. As time advances and subject data is updated, the scatter points (subjects) move around in the space, defining low-dimensional trajectories.

Given that we are dealing with a multidimensional, nonlinear mapping, similarity-ordering in the MKL space can follow complex patterns. For the sake of example, we illustrate cases where clinical variables appear highly ordered along a single dimension of the MKL space. As criterion for selection, we used the Pearson correlation coefficient between the (MKL) space dimension and (input) feature. The values for all such pairs are available in Figure 3.12 of Appendix 3.C. Herein, we discuss the highest correlation cases. For instance, the first dimension of the obtained space (Figure 3.4, top row) appears to be highly correlated with cervical dilatation, duration of contractions and, inversely, with the time between contractions. Thus, in this dimension, women in similar stages of labour are closely positioned, with the leftmost and rightmost regions of the scatter plots mostly populated with women that, at admission-time, were in earlier and later labour stages, respectively. It is then expected that subjects move towards the right, in the scatter plot, as labour advances. This trend is illustrated by Figures 3.5a and 3.5b. Figure 3.5a overlays the trajectories defined by some of the subjects of the training set

on the admission-time scatter plot of dimension 1 vs. dimension 2. Each sequence of connected triangles corresponds to the trajectory of one individual, with each triangle corresponding to a follow-up and colored by the follow-up timing normalized by delivery timing. An heterogeneity in initial positioning (i.e. admission-time labour stage) is observed. Nonetheless, as expected, all individuals define a rightwards trajectory as labour progresses. In Figure 3.5b, the $E_0^1(t)$ ($\pm\sigma_0^1(t)$) curve is plotted for a subject whose initial projection lies on the leftmost region of the scatter plot. As expected, with time, projection values in dimension 1 increase. An initially larger slope gradually decreases, a pattern that is explained by the fact that in the first few hours both slower and faster deliveries are weighing in on the curve estimation, whereas for later timings only slower deliveries are, pushing the mean curve down.

The position in the lower-dimensional space is not only dictated by dynamic labour variables. In the bottom row of Figure 3.4, we can observe that position in dimension 2 has some correlation with the country variable. Interestingly, it also correlates with cervix consistency, suggesting some association between country and admission-time assessment of cervix consistency. The leftmost scatter plot suggests that experiencing emotional distress is translated into a downwards displacement in dimension 6.

Figures 3.4 and 3.5 show some intuitive examples of similarity-ordering and clinical interpretability in the MKL space. However, as previously mentioned, similarity-ordering does not always happen in such obvious ways for all features/dimensions.

This interpretability of the MKL space can facilitate the identification of patterns regarding the occurrence of target events. For example, in Figure 3.6, analogous scatter plots are generated, this time colored by the (non)occurrence of our outcomes of interest, CS and BO. Figure 3.6-left seems to showcase a higher density of CS cases in subjects on the leftmost region, which we have seen to correspond to earlier-stage labours. This trend is confirmed in Figure 3.6-right. In the case of BO, there is a less evident pattern. Figure 3.11 in

Appendix 3.B extends this analysis to the practice of amniotomy and labour augmentation, with the resulting patterns suggesting some correlation between the incidence of these interventions and subject positioning along dimension 2 (Figure 3.11-right). Given the correlation of dimension 2 with country observed in Figure 3.4, this pattern suggests a higher incidence of both interventions within Nigeria's facilities.

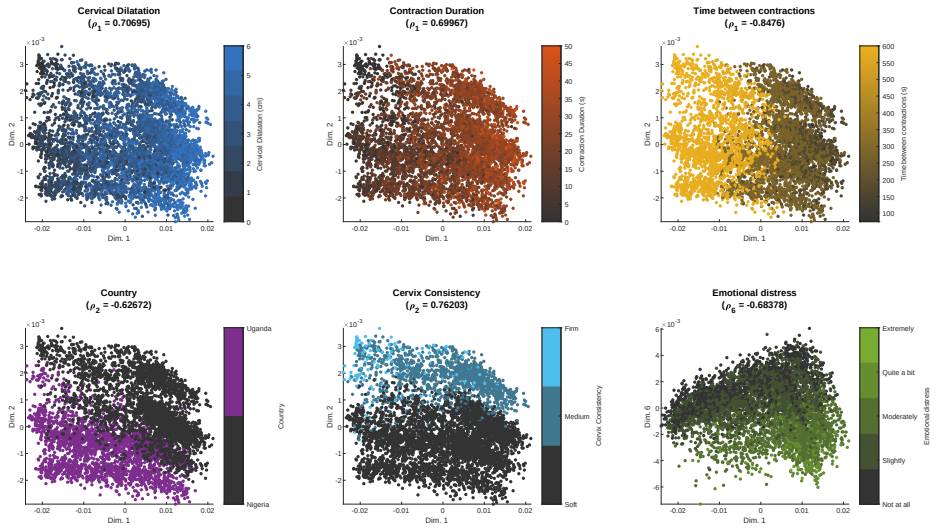
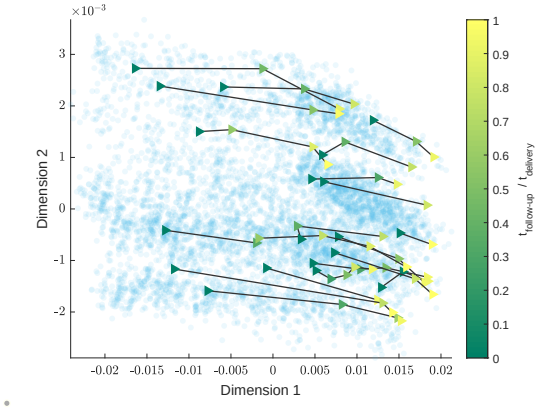
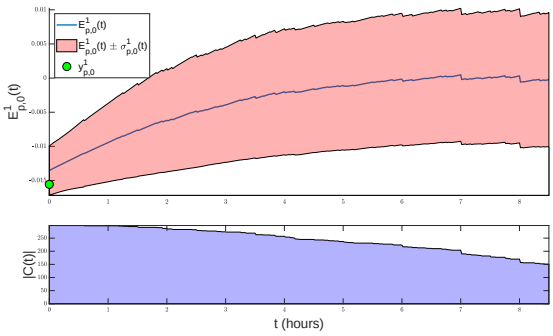


Figure 3.4: Similarity-based spatial ordering in the MKL space with the SELMA dataset. Each plot corresponds to the projections of the samples used to learn the MKL model, color-coded by a specific clinical variable that highly correlates with one of the dimensions of the MKL space. ρ_d = Pearson correlation coefficient between dimension d and the clinical variable.



(a)



(b)

Figure 3.5: Interpreting trajectories in the MKL space. (a) Examples of trajectories defined by training set subjects in the first 2 dimensions of the MKL space. Each sequence of connected triangles corresponds to the trajectory of one subject; the triangles correspond to follow-ups and are colored by respective follow-up timing normalized by delivery timing (taking admission-time as reference). (b) Example of initial estimate of expected progress along the first dimension for a subject with a low $y^1_{p,0}$ value. Top – $E^1_{p,0}(t)$, cropped at the timing where $|\mathbb{C}(t)|$ is halved. Bottom – count of the number of peers with uncomplicated labours used to estimate $E^1_{p,0}(t)$, $|\mathbb{C}(t)|$.

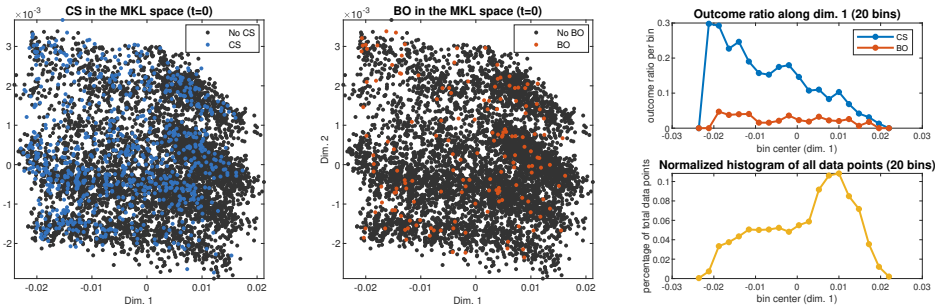


Figure 3.6: Spatial distribution of outcomes of interest in the admission-time MKL space. Right: CS and BO rates of occurrence throughout dimension 1, obtained by dividing scatter points in 20 bins along dimension 1 and computing each bin's occurrence rate.

3.3.2. Evaluation

The lack of a standardized reference for labour monitoring and decision making can lead to heterogeneous practice. To explore practice differences within our dataset, additionally to evaluating the framework from a global perspective, we also engage in subgroup-level analysis.

Global performance

Table 3.1 contains the results of the CS prediction experiments for the complete training and testing populations (columns styled in bold). Regarding the cross-validation stage, $\{v_k^{CS}\}, k \in \{\pi, \pi z, \pi zt\}$, showed similar performances, with $v_{\pi zt}^{CS}$ performing marginally better. The AUC values, ranging from 0.746 to 0.767, suggest a reasonably good predictive power. In what regards the other performance metrics, it is observed that, with the selected cut-offs, the framework evaluation predictors largely outperformed the classical alert and action lines, achieving a much better trade-off between metrics related to the positive (SE, PPV) and negative (SP, NPV) class, with $v_{\pi zt}^{CS}$ achieving SE and SP ≈ 0.7, ≈ 0.26 and NPV ≈ 0.94. The alert and action lines present relatively good specificity, at the expense of poor sensitivity. When applying the learned cut-offs to the testing set, performances did not significantly change, suggesting a good generalizability.

Table 3.2 shows the results of the BO prediction experiments. Again, $\{v_k^{BO}\}, k \in \{\pi, \pi z, \pi zt\}$, showed similar performances, this time with v_{π}^{BO} performing slightly better. Compared to the CS prediction experiments, predictive performances were significantly lower. Despite AUC values (ranging between 0.561 and 0.594) being only slightly better than that of the random classifier, results of random permutation tests suggest that this improvement is statistically significant (p-values ≤ 0.0008). With the selected cut-offs, the defined predictors again achieved a better trade-off between metrics related to the positive and negative class, with SE and SP above 0.56, PPV

Table 3.1: CS prediction results. n = sample size; n_{CS} = number of positive cases; Th = threshold/cut-off; SE = sensitivity; SP = specificity; PPV = positive predictive value; NPV = negative predictive value; AUC = area under the receiver operating characteristic; p-value = fraction of random permutation tests for which $AUC \geq AUC_{observed}$ (total of 10000).

Train ($n = 6349$; $n_{CS} = 817$)						
	Th	**SE**	**SP**	**PPV**	**NPV**	**AUC (p-value)**
Alert line	-	0.540	0.728	0.227	0.915	-
Action line	-	0.290	0.889	0.278	0.894	-
v_{π}^{CS}	0.221	0.699	0.700	0.256	0.940	0.763 (< 0.0001)
$v_{\pi z}^{CS}$	0.422	0.683	0.684	0.242	0.936	0.746 (< 0.0001)
$v_{\pi zt}^{CS}$	2.038	0.706	0.707	0.263	0.942	0.767 (< 0.0001)
Test ($n = 2121$; $n_{CS} = 279$)						
	Th	**SE**	**SP**	**PPV**	**NPV**	**AUC (p-value)**
Alert line	-	0.548	0.731	0.236	0.914	-
Action line	-	0.290	0.891	0.288	0.892	-
v_{π}^{CS}	0.221	0.674	0.696	0.251	0.934	-
$v_{\pi z}^{CS}$	0.422	0.659	0.712	0.258	0.932	-
$v_{\pi zt}^{CS}$	2.038	0.703	0.712	0.270	0.941	-

≈ 0.03 and NPV ≈ 0.98, whereas the alert and action lines again favored specificity over sensitivity. When applying the learned cut-offs to the testing set, the SE-SP balance decreased in the cases of v_{π}^{BO} and $v_{\pi z}^{BO}$. On the other hand, $v_{\pi zt}^{BO}$ showed good generalizability.

Subgroup performances I: the four-centimeter threshold

The results of the first subgroup analysis are shown in Table 3.3. Note that (as expected given the patterns observed in Figure 3.6) the CS rates are significantly larger for the "less than 4 cm" subgroups, which might have a direct effect on PPV and NPV. Before looking at specific cut-offs, there is already an evident difference between the average AUC values of the two groups (with $AUC^{\uparrow}$ and $AUC^{\downarrow}$ reaching a maximum of 0.67 and 0.81, respectively). This gap reflects on the remaining performance metrics, obtained with the recomputed cut-offs. Having the global-level experiments as reference, in the case of the "less than 4 cm" subgroup, cut-off values adapt by increasing,

Table 3.2: BO prediction results. n = sample size; n_{BO} = number of positive cases; Th = threshold/cut-off; SE = sensitivity; SP = specificity; PPV = positive predictive value; NPV = negative predictive value; AUC = area under the receiver operating characteristic; p-value = fraction of random permutation tests for which $AUC \geq AUC_{observed}$ (total of 10000).

Train ($n = 6349$; $n_{BO} = 155$)						
	Th	**SE**	**SP**	**PPV**	**NPV**	**AUC (p-value)**
Alert line	-	0.419	0.697	0.033	0.980	-
Action line	-	0.174	0.867	0.032	0.977	-
v_{π}^{BO}	0.036	0.594	0.594	0.035	0.983	0.612 (< 0.0001)
$v_{\pi z}^{BO}$	0.069	0.561	0.567	0.031	0.981	0.581 (0.0008)
$v_{\pi zt}^{BO}$	0.283	0.568	0.573	0.032	0.981	0.595 (< 0.0001)
Test ($n = 2121$; $n_{BO} = 44$)						
	Th	**SE**	**SP**	**PPV**	**NPV**	**AUC (p-value)**
Alert line	-	0.455	0.698	0.031	0.984	-
Action line	-	0.205	0.869	0.032	0.981	-
v_{π}^{BO}	0.036	0.523	0.635	0.029	0.984	-
$v_{\pi z}^{BO}$	0.069	0.500	0.605	0.026	0.983	-
$v_{\pi zt}^{BO}$	0.283	0.568	0.557	0.026	0.984	-

and the opposite happens in the complementary subgroup. Figure 3.7 shows how performances are being optimized for both subgroups by using adaptive cut-offs, as opposed to a globally estimated one. Performances of the alert and action lines were again significantly inferior to those of the framework evaluation predictors.

Figure 3.8 compares the performance of our framework evaluation predictor $v_{\pi zt}^{CS}$ with those of the admission-time and earliest interval (0-2 h after onset of 4 cm of cervical dilatation) models by Souza et al. [Souza et al., 2019] (referred to as Model 1 and Model 2 in said publication, respectively), demonstrating that we achieve comparable performances to those of their prediction models.

When doing the same evaluation for the BO prediction task, performances were poor for both subgroups (Table 3.4).

Table 3.3: CS prediction results for "less than 4 cm" and "4 cm and over" subgroups ("↓" and "↑" superscripts, respectively). n = sample size; n_{CS} = number of positive cases; Th = threshold/cut-off; SE = sensitivity; SP= specificity; PPV = positive predictive value; NPV = negative predictive value; AUC = area under the receiver operating characteristic.

Train ($n^{\downarrow} = 2039$; $n^{\downarrow}_{CS} = 396$; $n^{\uparrow} = 4310$; $n^{\uparrow}_{CS} = 421$)												
	$Th^{\downarrow}$	$Th^{\uparrow}$	$SE^{\downarrow}$	$SE^{\uparrow}$	$SP^{\downarrow}$	$SP^{\uparrow}$	$PPV^{\downarrow}$	$PPV^{\uparrow}$	$NPV^{\downarrow}$	$NPV^{\uparrow}$	$AUC^{\downarrow}$	$AUC^{\uparrow}$
Alert line	-	-	0.652	0.435	0.502	0.824	0.240	0.211	0.857	0.931	-	-
Action line	-	-	0.434	0.154	0.726	0.958	0.276	0.284	0.842	0.913	-	-
v_{π}^{CS}	0.266	0.179	0.614	0.715	0.616	0.715	0.278	0.214	0.869	0.959	0.670	0.784
$v_{\pi z}^{CS}$	0.583	0.334	0.596	0.713	0.598	0.713	0.263	0.212	0.860	0.958	0.637	0.775
$v_{\pi zt}^{CS}$	4.661	1.253	0.606	0.739	0.609	0.740	0.272	0.235	0.865	0.963	0.646	0.813
Test ($n^{\downarrow} = 708$; $n^{\downarrow}_{CS} = 139$; $n^{\uparrow} = 1413$; $n^{\uparrow}_{CS} = 140$)												
	$Th^{\downarrow}$	$Th^{\uparrow}$	$SE^{\downarrow}$	$SE^{\uparrow}$	$SP^{\downarrow}$	$SP^{\uparrow}$	$PPV^{\downarrow}$	$PPV^{\uparrow}$	$NPV^{\downarrow}$	$NPV^{\uparrow}$	$AUC^{\downarrow}$	$AUC^{\uparrow}$
Alert line	-	-	0.655	0.443	0.524	0.824	0.251	0.217	0.861	0.931	-	-
Action line	-	-	0.439	0.143	0.754	0.953	0.303	0.250	0.846	0.910	-	-
v_{π}^{CS}	0.266	0.179	0.626	0.750	0.643	0.727	0.300	0.232	0.876	0.964	-	-
$v_{\pi z}^{CS}$	0.583	0.334	0.525	0.700	0.663	0.727	0.275	0.220	0.851	0.957	-	-
$v_{\pi zt}^{CS}$	4.661	1.253	0.576	0.743	0.634	0.761	0.278	0.255	0.860	0.964	-	-

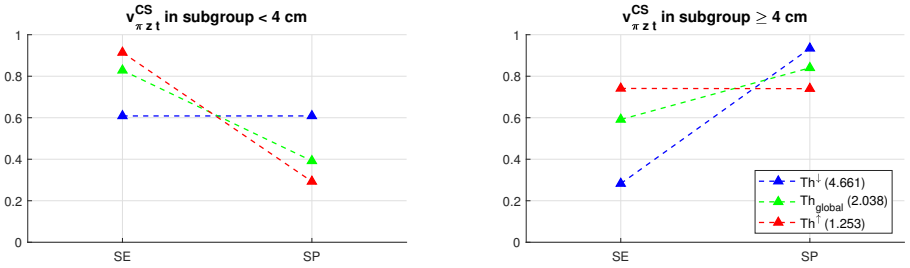


Figure 3.7: Illustration of cut-off value adaptation to optimize performances at the subgroup level. SE and SP pair for the predictor $v^{CS}_{\pi zt}$ in the “less than 4 cm” (left) and “4 cm and over” (right) subgroups, when the estimated global and subgroup cut-offs are used.

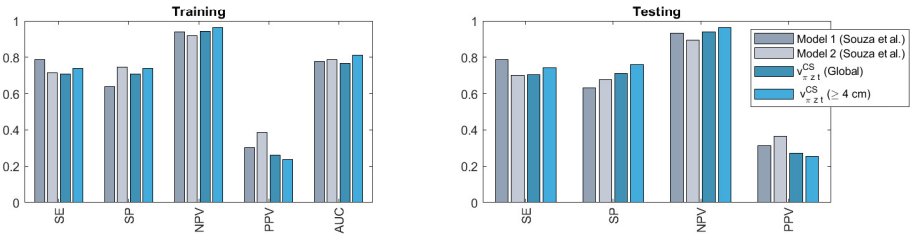


Figure 3.8: Comparison of obtained performances with those of admission-time (Model 1) and earliest interval (Model 2) models by Souza et al. [Souza et al., 2019].

Table 3.4: BO prediction results for "less than 4 cm" and "4 cm and over" subgroups ("↓" and "↑" superscripts, respectively). n = sample size; n_{BO} = number of positive cases; Th = threshold/cut-off; SE = sensitivity; SP = specificity; PPV = positive predictive value; NPV = negative predictive value; AUC = area under the receiver operating characteristic.

Train ($n^{\downarrow} = 2039$; $n^{\downarrow}_{BO} = 73$; $n^{\uparrow} = 4310$; $n^{\uparrow}_{BO} = 82$)												
	$Th^{\downarrow}$	$Th^{\uparrow}$	$SE^{\downarrow}$	$SE^{\uparrow}$	$SP^{\downarrow}$	$SP^{\uparrow}$	$PPV^{\downarrow}$	$PPV^{\uparrow}$	$NPV^{\downarrow}$	$NPV^{\uparrow}$	$AUC^{\downarrow}$	$AUC^{\uparrow}$
Alert line	-	-	0.548	0.305	0.473	0.801	0.037	0.029	0.966	0.983	-	-
Action line	-	-	0.233	0.122	0.692	0.948	0.027	0.044	0.960	0.982	-	-
v_{π}^{BO}	0.043	0.033	0.493	0.622	0.494	0.635	0.035	0.032	0.963	0.989	0.502	0.662
$v_{\pi z}^{BO}$	0.094	0.055	0.479	0.549	0.486	0.552	0.033	0.023	0.962	0.984	0.482	0.603
$v_{\pi zt}^{BO}$	0.651	0.177	0.479	0.573	0.490	0.573	0.034	0.025	0.962	0.986	0.482	0.625
Test ($n^{\downarrow} = 708$; $n^{\downarrow}_{BO} = 17$; $n^{\uparrow} = 1413$; $n^{\uparrow}_{BO} = 27$)												
	$Th^{\downarrow}$	$Th^{\uparrow}$	$SE^{\downarrow}$	$SE^{\uparrow}$	$SP^{\downarrow}$	$SP^{\uparrow}$	$PPV^{\downarrow}$	$PPV^{\uparrow}$	$NPV^{\downarrow}$	$NPV^{\uparrow}$	$AUC^{\downarrow}$	$AUC^{\uparrow}$
Alert line	-	-	0.706	0.296	0.493	0.799	0.033	0.028	0.986	0.983	-	-
Action line	-	-	0.529	0	0.722	0.942	0.045	0.000	0.984	0.980	-	-
v_{π}^{BO}	0.043	0.033	0.529	0.407	0.583	0.686	0.030	0.025	0.981	0.983	-	-
$v_{\pi z}^{BO}$	0.094	0.055	0.529	0.444	0.556	0.595	0.028	0.021	0.980	0.982	-	-
$v_{\pi zt}^{BO}$	0.651	0.177	0.706	0.444	0.507	0.584	0.034	0.020	0.986	0.982	-	-

Subgroup performances II: finer granularity

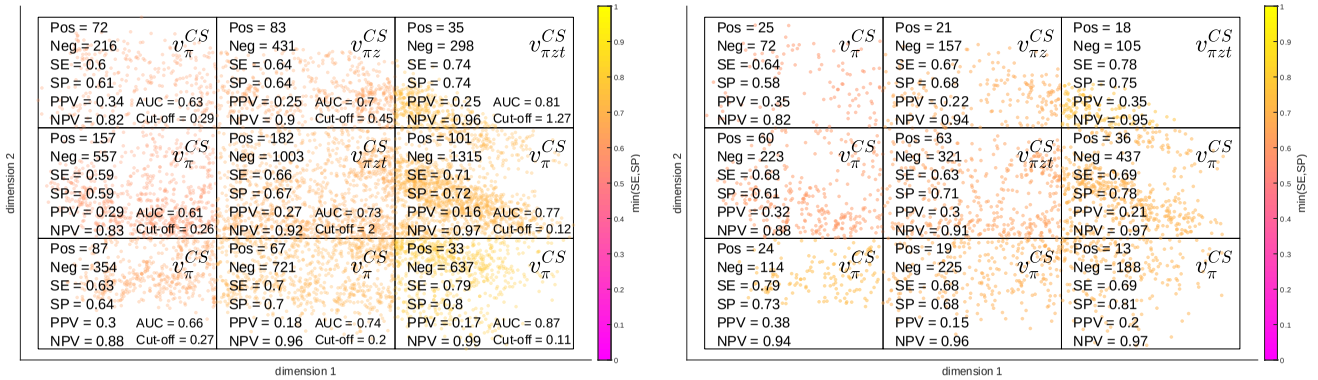
In the previous analyses we observed that performances can be improved by estimating subgroup cut-offs for our predictors. So far, we have considered only two subgroups: cervical dilatation less than/greater or equal than 4 cm upon arrival. Herein, we increase the granularity of the analysis, and look at smaller subgroups of the population. The subgroups are defined as equally sized regions of the admission-time MKL space (dimension 1 vs. dimension 2). Figures 3.9 and 3.10 illustrate the partitioning of the space in said regions for the training (left) and testing (right) set subjects. In the top row plots, the predictor among $\{v_k^e\}, k \in \{\pi, \pi z, \pi zt\}$, with the best performance in testing (measured as the maximum value for $min(SE, SP)$) is identified for each region, along with the selected cut-off and corresponding performance metrics (Figures 3.9a and 3.10a). In the bottom row plots, the same process is repeated for the alert and action lines (Figures 3.9c and 3.10c). When each regional cut-off is applied to the corresponding partition of the testing set, the performances are those highlighted in Figures 3.9b, 3.10b, 3.9d and 3.10d. All scatter plots are colored by the regional minimum between SE and SP.

The color patterns in Figures 3.9 and 3.10 suggest that, consistently with the results of the global analysis, (1) the framework evaluation predictors present overall superior performances to the partograph predictors, and (2) the CS predictive performances are significantly higher than those regarding BO.

Figure 3.9a reveals a gradient in performance that is roughly organized in rightward orientation, suggesting that the prediction of CS is more effective for subgroups corresponding to later labour stages at admission time. On the other hand, the optimal cut-off values decrease as we move to the right. These effects are consistent with the results of the previous analysis based on the admission-time 4 cm threshold. The v_π^{CS} predictor appears 6 out of 9 times as that with the best performance in testing, among the three candidates. In most

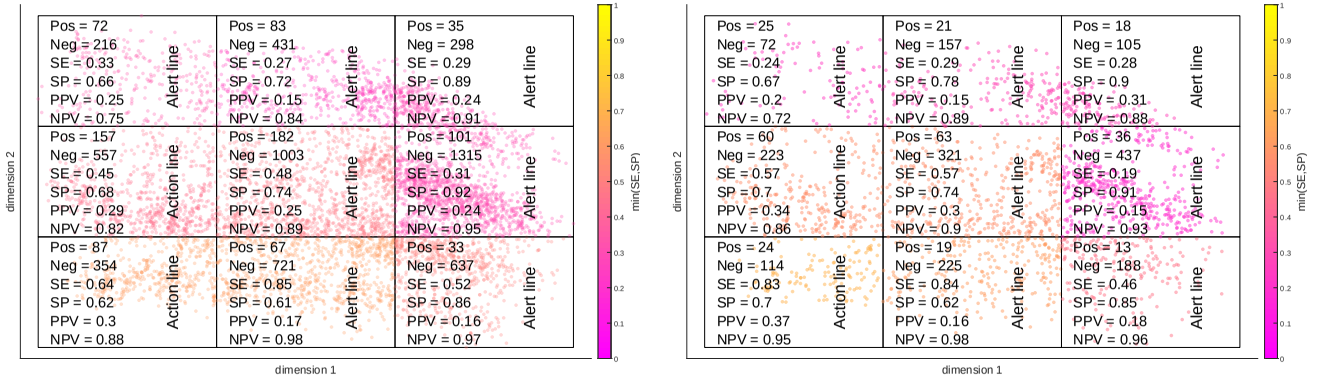
cases, performances in the testing set (Figure 3.9b) are comparable to those in the training set. The regional performances of the alert and action lines (Figures 3.9c and 3.9d) range from comparable to significantly worse than those of the framework evaluation predictors, depending on the subject subgroup at hand. It is observed that they perform best in partitions where Uganda is the dominant country (see Figure 3.4).

When it comes to BO prediction, regional analysis was done at a coarser level due to the scarcity of positive cases. Regional performances of the framework evaluation predictors (Figures 3.10a and 3.10b) were more or less in line with the global counterpart (Table 3.2), except for two partitions where performance levels were significantly worse. In the 4 remaining partitions, the value of $min(SE, SP)$ was above 0.56 in both training and testing, with a prevalence of the $v_{\pi zt}^{BO}$ predictor over the two other candidates. There is no apparent organized spatial pattern in performance. The regional performances of the partograph predictors (Figures 3.10c and 3.10d) were either comparable the framework evaluation predictors, or (most often) significantly worse than the framework evaluation predictors – only in 1 out of 6 partitions both SE and SP values surpassed 0.5 in both training and testing; moreover, in 4 out of 5 partitions there was a very poor SE-SP trade-off (again in favour of specificity).



(a) Framework evaluation predictors, training (cross-validation) set.

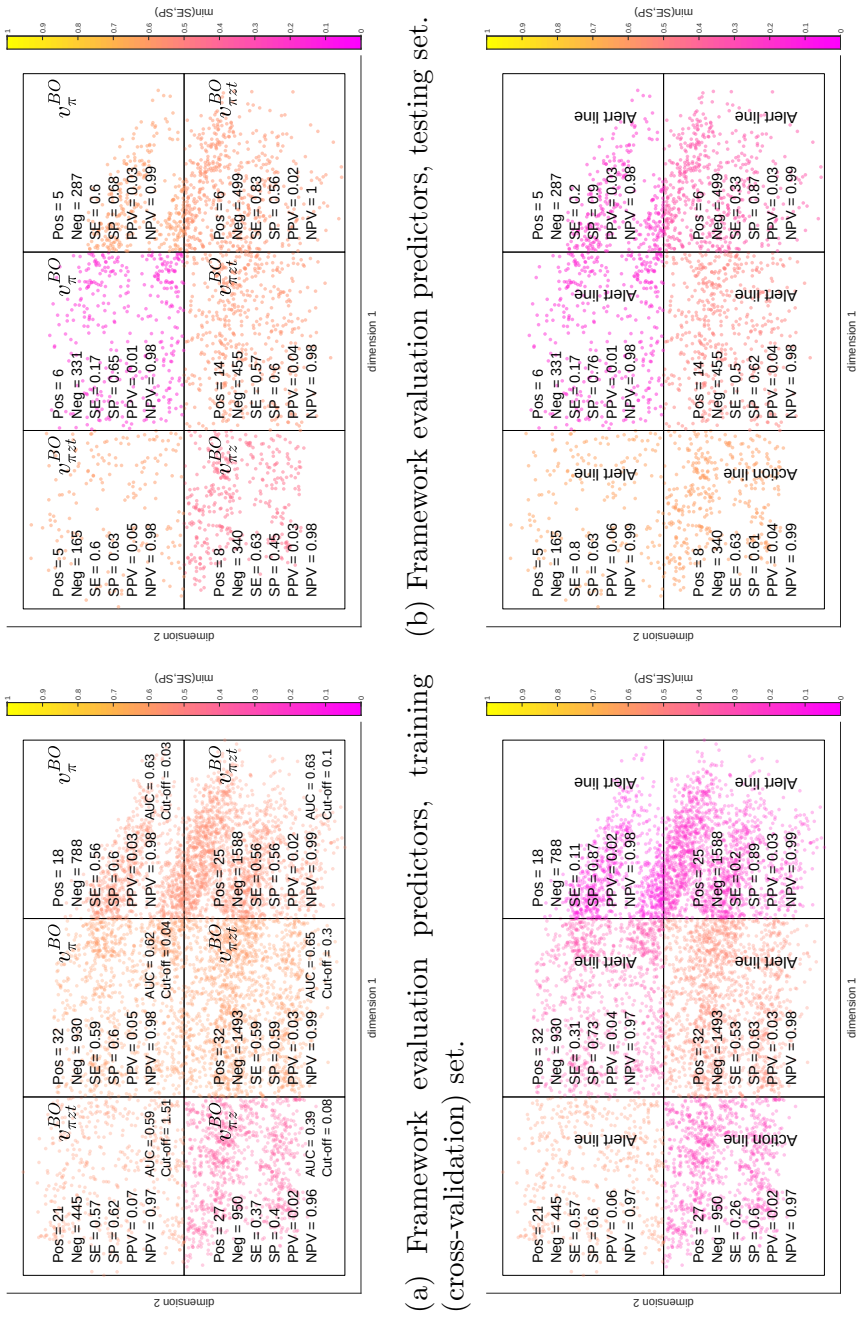
(b) Framework evaluation predictors, testing set.



(c) Partograph predictors, training (cross-validation) set.

(d) Partograph predictors, testing set.

Figure 3.9: Regional CS predictive performances in the training (left) and testing (right) sets, for the framework evaluation predictors, $\{v_k^{CS}\}, k \in \{\pi, \pi z, \pi zt\}$, (top), and classical partograph predictors (bottom). Scatter points colored by $min(SE, SP)$.



(a) Framework evaluation predictors, training (cross-validation) set.

(b) Framework evaluation predictors, testing set.

(c) Partograph predictors, training (cross-validation) set.

(d) Partograph predictors, testing set.

Figure 3.10: Regional BO predictive performances in the training (left) and testing (right) sets, for the framework evaluation predictors, $\{v_k^{BO}\}, k \in \{\pi, \pi z, \pi zt\}$, (top) and classical partograph predictors (bottom). Scatter points colored by $min(SE, SP)$.

3.4. Discussion

We presented a novel approach for personalized temporal labour monitoring and decision support. The proposed framework dynamically identifies peers (from a cohort study) of the mother to be monitored, based on similarity of the different input variables. Peer labour progress, intervention and outcome data are then used to evaluate divergence from ideal progress and to provide estimates of risk of adverse outcome or intervention recommendations. The proposed labour monitoring framework thus addresses the main limitations of the partograph (univariate, one-fits-all) approach and enriches it with reference practice data. Besides describing the proposed algorithm, we evaluated its performance (when used in a simple decision support system) to predict adverse fetal/maternal labour outcome and caesarean section. For validation, we compared performances with the current state of the art: the current recommended monitoring tool (the partograph) and some of the most recent and best performing CS prediction models, presented by Souza et al. [Souza et al., 2019]. Additionally, in order to explore the potential of the framework in the identification and learning of different practices, we performed subgroup analyses.

While being a machine learning approach, the proposed framework uses a fully interpretable paradigm that relies on non-supervised information similarity to calculate distances of subjects based on a complex and comprehensive data. This similarity is used to position subjects within a lower-dimensional space that can easily be interpreted in regards to the input data. To illustrate this clinical interpretability, we calculated the Pearson correlation of the resulting distribution with the input information and provided a labeled visualisation, showing which variables predominately identify peers of a given mother. This visualisation provides meaningful insight into the cohort data, and their interactions, used for training; for example, in case of the SELMA study, it highlights the different practice in the participating countries. Additionally, when using temporally varying

dynamic information, labour progress can be interpreted from the trajectory within the lower-dimensional space, and different subjects can be compared amongst each other or to available cohort data. An additional advantage of the proposed framework is that it naturally follows the paradigm of the parthograph while addressing some of its shortcomings. The underlying idea of the partograph is to visually provide feedback on the dynamics of an extended set of variables and compare this continuously to progress that is perceived as leading to desired outcome. When this comparison deviates from normality (i.e. crosses the alert or action line), an indication for intensified monitoring or corrective actions is provided. Our approach uses a similar paradigm but additionally personalises and continuously updates the information an individual is compared to, in order to make it more pertinent for that individual, and it allows using knowledge from cohort or clinical studies to suggest the likelihood of a specific intervention.

The performance evaluation results show that our framework significantly outperformed the partograph's alert and action lines in the prediction of both CS and BO (Tables 3.1 and 3.2; Figures 3.9 and 3.10). The predictive performances were significantly higher for CS. The lower rates for predicting BO were expected given the very low occurrence (only about 2%) as well as the fact that interventions such as CS are performed exactly to prevent BO when this is suspected during labour progress. The diversity in etiologies, as well as the inclusion of postpartum events, when available monitoring data ends at delivery, further complicate the prediction task.

Subgroup analysis allowed us to detect differences in CS practice among groups of individuals with different admission-time characteristics, and adapt cut-offs to maximize subgroup-level performances. This type of analysis let us understand, for example, that CS practice was more aligned with the alert and action lines in Ugandan than in Nigerian facilities; nonetheless, our approach was able to "learn" both countries' practices with comparable performances. We observed that CS was significantly more prevalent among women with lower

admission-time dilatations, a result that is consistent with evidence from previous observational studies [Holmes et al., 2001, Neal et al., 2014, Bailit et al., 2005, Mikolajczyk et al., 2016, Chuma et al., 2014]. However, the achieved local performances with adapted cut-offs were far superior for late-admission subgroups. In other words, CSs were performed in significantly larger amounts, but seemingly in a less consistent/predictable manner. Overall, subgroup predictive analysis demonstrates high potential in the identification of practice differences/biases and in the subgroup-level optimization of predictive performance. In the case of BO, the previously enumerated challenging aspects of BO prediction are the same that limit our ability to discover subgroup patterns.

It was observed that combining cohort-based occurrence estimates with distance from normality and timing information added predictive value in multiple occasions, in both CS and BO prediction tasks. Using a more complex combination of predictors or calculation of distance might improve the results.

Previous approaches to decision support in terms of CS consisted mostly of supervised training models to learn study-specific practice from a subset of the study and testing predictive performance on a different subset of the same study. It is thus important to clarify that good predictive performances suggest that they succeed at learning "what was done", which does not always align with "what should have been done". This is specifically the case for CS where the clinical decision is based on a combination of true clinical need and local practice/preference. This requires that risk estimates obtained from these models, albeit useful, should be handled with caution and ideally always as a complement to other risk estimates. Because there is no ground truth data on "what should have been done", "what was done" is commonly used to evaluate systems, which is why we also evaluated our framework for this problem. Our approach differs from others in that it is not explicitly developed and trained to maximize performance in this one (and only) specific task. We rather include study-specific chance estimates as one of the elements to weigh in on

risk assessment, which happens to be the component that is possible to evaluate with existing data. Additionally we provide an independent quantification of how the progress of an individual deviates from peers that went through uneventful labours. The value of this can only be fully evaluated in a wider prognostic evaluation.

Recently Souza et al. [Souza et al., 2019] presented an optimised supervised logistic regression model for the prediction of CS. Instead of one single model, they suggested to use a combination of a labour admission model, interval models, and a maximum score model. Our framework performed comparably to the admission model (Figure 3.8). Their interval models require having an assessment made at a cervical dilatation of 4 cm, which inherently limits their clinical use to women with cervical dilatations of 4 cm and under upon admission. In the case of SELMA dataset, that information was only present in less than 30% of the individuals. Our system (useful in everyone irrespective of admission dilatation or labour phase) showed similar performance for their proposed [0-2 hours]-interval model (Figure 3.8), while at the same time providing a much more intuitive approach towards decision making. The reported improved performance in the later-interval models is difficult to compare with our results given that they can only be used in 15% (for the 4 hours-model) or 5% (for the 6 hours model) of individuals. Their maximum score model does outperform our predictors. However, this model is based on a post-hoc analysis of the data throughout labour progression, thus making it impossible to use for decision support in a real-time scenario where at each time point a decision regarding a possible intervention needs to be taken. Additionally, it should be recalled that we are comparing the performance of a minimal implementation of our proposed framework, assessed based on simple individual predictors, with that of multivariate prediction models specifically trained for this prediction task.

3.5. Limitations

The main limitations of the current paper are related to: (1) the specific implementation choices for each stage of the framework – current methodologies behind the different steps can be replaced with more suited/sophisticated techniques; and (2) the evaluation scheme – the naively defined predictors and evaluation scheme are unlikely those which maximize performance. Nonetheless, the results clearly show the power of the novel proposed paradigm.

The fact that the learning of the MKL model is completely unsupervised gives us limited control over which features will be dominant in the ordering of data in the space. In some cases, this might be seen as a limitation. For example, features with very rare occurrences of abnormal values, even if correlated with outcome, might end up underrepresented. One can gain some control over the obtained space, by relaxing the "unsupervised" constraint. For example, the model could be learned supervised by outcome, thus favoring features that are discriminatory with regard to that outcome. Another alternative would be to have prior clinical knowledge on which features are more important in contributing to the model estimation, a solution that, although not completely unsupervised, would be less dependent on the outcome of a particular study.

On the other hand, the main contribution of the current paper is the design of a monitoring framework that is flexible both in implementation and in application (i.e. easily translatable to other clinical monitoring and decision support problems).

3.6. Conclusions

We proposed a labour monitoring framework that addresses some of the main limitations of the current reference tool, the partograph, as well as of logistic regression models optimised for predicting certain events.

A similarity-based dimensionality reduction step enables a simplified interpretable representation of high-dimensional data. This representation might help, first, in the identification of complex patterns of interaction among clinical variables that are hard to perceive by the classical visualization of the partograph. Then, this representation is key for our formulation of a peer-based personalization of labour monitoring. Under the premise of a precomputed database of projected labour data with known interventions and outcomes, the framework dynamically evaluates "normality" in labour progress and also provides insight on what would be study practice in similar scenarios.

Experiments with the SELMA study illustrate the clinical interpretability of the framework and its superiority compared to the partograph's alert and action lines in the prediction of clinically relevant events. Additionally, it is shown that, with a minimal implementation and an evaluation based on simple and intuitive descriptors, it performs comparably to state-of-the-art multivariate prediction models, while tackling some of their limitations in terms of integration in a clinical environment.

Overall, we believe that the current paper showcases the proposed framework as a promising alternative way of looking at the problem of labour monitoring and allows extension to any decision-making or prognosis task in the setting of clinical reality where subjects show a very heterogeneous initial presentation, are monitored in non-standardised intervals and are evaluated towards complex outcomes.

Appendices

3.A. Static and Dynamic Features

Table 3.5: Admission-only / static features.

	NAME	NOTES
1	Country code	Country code: Uganda/Nigeria (1/0)
2	**Ethnicity**	Ethnicity: [NIGERIA] 1 - Ibo; 2 - Yoruba; 3 - Hausa; 4 - Fulani; 5 - TIV; 6 - Kanuri; 7 - Other Nigerian; 8 - Non Nigerian; [UGANDA] 9 - Muganda/Musoga/Mugisu; 10 - Munyakore/Mukiga/Munyoro/Mutoro; 11 - Acholi/Langi/Alur; 12 - Iteso/Karamojong; 13 - Lugbara/Madi; 14 - Other Ugandan; 15 - Non-Ugandan
3	Facility code	1-13
4	Age	years
5	Height	cm
6	Foot length	cm
7	Current weight	kg
8	Marital status	Marital status: 0 - Single / Separated / Divorced / Widowed; 1 - Married / Cohabitating

9	**Education level**	Education level: 0 - No education; 1 - Other (e.g. Quranic / Nomadic education only; 2 - Pre-primary education; 3 - Incomplete primary education; 4 - Complete primary education; 5 - Incomplete secondary education; 6 - Complete secondary education; 7 - Incomplete post-secondary/tertiary education; 8 - Complete post-secondary/tertiary education)
10	Gainful occupation	Gainful occupation: 0 - No; 1 - Yes
11	Parity	Number of previous births
12	**Previous abortions or stillbirths**	Previous abortions or stillbirths: 0 - No; 1 - Yes
13	**Previous uterine surgery**	Previous uterine surgery (includes previous c-sections or other uterine surgeries): 0 - None; 1 - One; 2 - More than one
14	Best estimate of gestation	weeks
15	**Mode of labour onset and referral (or not) from another health facility**	Mode of labour onset and referral (or not) from another health facility: 0 - spontaneous onset, not referred from another facility; 1 - induced, not referred; 2 - spontaneous, referred; 3 - induced, referred
16	**Fetal movements in the last 2h**	Fetal movements in the last 2h: 0 - reduced or absent; 1 - no changes/increased
17	Preterm rupture of membranes	Preterm rupture of membranes: 0 - No; 1 - Yes
18	**Obstetric haemorrhage**	Placenta praevia, accreata increta percreta, placentae abruption or other obstetric haemorrhage: 0 - No; 1 - Yes
19	**Pre-eclampsia or eclampsia**	Pre-eclampsia or eclampsia: 0 - No; 1 - Yes

20	Cervix effacement	Cervix effacement: 0 - Thick (less than 30% effaced); 1 - Medium (up to 50% effaced); 2 - Thin (up to 80% effaced); 3 - Very thin / paper-thin (more than 80% effaced)
21	Cervix position	Cervix position: 0 - Anterior; 1 - Central; 2 - Posterior
22	Cervix consistency	Cervix consistency: 0 - Soft; 1 - Medium; 2 - Firm
23	Symphysis fundal height	cm
24	Sacral promontory reached	Sacral promontory reached: 0 - No; 1 - Yes; 2 - Not assessed
25	Ischial spines prominent	Ischial spines prominent: 0 - No; 1 - Yes; 2 - Not assessed
26	Pubic angle admits less than two fingers	Pubic angle admits less than two fingers: 0 - No; 1 - Yes; 2 - Not assessed
27	**Cardiovascular condition**	Chronic hypertension, heart_disease, obesity, or chronic_anaemia: 0 - No; 1 - Yes
28	**Immunity condition**	HIV or AIDS: 0 - No; 1 - Yes
29	**Diabetes**	Diabetes or gestational diabetes: 0 - No; 1 - Yes
30	**Renal condition**	Pyelonephritis or renal disease: 0 - No; 1 - Yes
31	Lung disease	Lung disease: 0 - No; 1 - Yes
32	Anaemia	Anaemia: 0 - No; 1 - Yes
33	**Other condition**	Other chronic disease, other pregnancy complications, malaria: 0 - No; 1 - Yes

Table 3.6: Follow-up / dynamic features.

	NAME	NOTES
1	Contraction ON time	Duration of uterine contractions (seconds)
2	**Contraction OFF time**	Time between contractions (seconds)
3	Cervical dilatation	cm
4	Maternal Heart Rate	bpm
5	Systolic Blood Pressure	mmHg
6	Diastolic Blood Pressure	mmHg
7	Axillary Temperature	°C
8	Amniotic membranes status	Amniotic membranes status: 0 - Intact; 1 - Ruptured without meconium; 2 - Ruptured with stale meconium; 3 - Ruptured with fresh meconium
9	Emotional status	Since the last assessment, how much the woman has been bothered by emotional problems such as fear, anxiety, depression, irritability, or sadness? 0 - Not at all; 1 - Slightly; 2 - Moderately; 3 - Quite a bit; 4 - Extremely
10	Labour pain	Since the last assessment, how much the woman has been bothered by labour pain? 0 - Not at all; 1 - Slightly; 2 - Moderately; 3 - Quite a bit; 4 - Extremely
11	Labour Companionship	Labour Companionship: 0 - No; 1 - Yes
12	Fetal Heart Rate	bpm
13	Fetal movements	Fetal movements observed/felt: 0 - No; 1 - Yes

14 Fetal presentation	Fetal presentation: 0 - Cephalic; 1 - Breech; 2 - Transverse lie / compound / other
15 Fetal station	Fetal station: 0 - Above ischial spine; 1 - At ischial spine; 2 - Below ischial spine
16 Position of fetal head	Position of fetal head: 0 - Occiput Anterior (includes right and left); 1 - Occiput transverse; 2 - Occiput posterior; 3 - Other
17 Caput Succedaneum	Caput Succedaneum: 0 - None; 1 - Mild; 2 - Moderate; 3 - Severe
18 Moulding	Moulding: 0 - None; 1 - First degree; 2 - Second degree; 3 - Third degree
19 Maternal position	Predominant maternal position between assessments: 0 - Upright, sitting, standing, walking, kneeing, squatting, all-4; 1 - Recumbent, semi-recumbent, lateral, supine

3.B. Other interventions in the MKL space

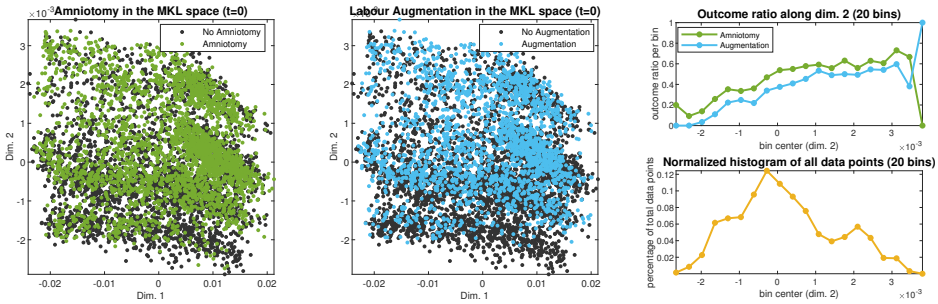


Figure 3.11: Spatial distribution of other interventions in the admission-time MKL space. Right: amniotomy and labour augmentation rates of occurrence throughout dimension 2, obtained by dividing scatter points in 20 bins along dimension 2 and computing each bin's occurrence rate.

3.C. Dimension-variable correlation coefficients

Figure 3.12: Pearson correlation coefficients of the 10 vs. 52 dimension-variable pairs, in the training data admission-time MKL space.

Chapter 4

BCN-SELMA: A SIMPLIFIED, EFFECTIVE, LABOUR MONITORING-TO-ACTION TOOL, BASED ON INTERPRETABLE MACHINE LEARNING

This chapter is adapted from: M. Nogueira, C. Yagüe, G. Piella, M. De Craene, S. Sanchez-Martinez, P. Martí, M. Bonet, O.T. Oladapo, B. Bijnens. BCN-SELMA: A Simplified, Effective, Labour Monitoring-to-Action tool, based on Interpretable Machine Learning. *In preparation.*

4.1. Background

4.1.1. SELMA

The World Health Organization (WHO)'s project for the development of a Simplified, Effective, Labour Monitoring-to-Action (SELMA) tool has as primary objectives [Souza et al., 2015]:

- "To identify the essential elements of intrapartum monitoring that trigger the decision to use interventions aimed at preventing poor labour outcomes";
- "To develop a simplified, monitoring-to-action algorithm for labour management";
- "To compare the diagnostic performance of SELMA and partograph algorithms as tools to identify women who are likely to develop poor labour-related outcomes".

To this end, a large database covering approximately 10.000 deliveries has been collected in a multicentric study, with a rich set of features being collected at first presentation, during different stages of labour and after delivery. Although the study provides a unique source of knowledge to address the anticipated objectives, there are two major obstacles to easily translate the study into a monitoring-to-action algorithm. Firstly, the incidence of adverse outcome in the study is low ($\approx 2\%$), which poses severe imbalance problems for any learning tool that uses this label for predictions. Secondly, and even more problematic for interpretation and learning, the study did not rigidly define what course of actions to be taken during labour using hard decision criteria/timings, but rather suggested to use best practice as implemented locally and recommended by the WHO. Although this obviously closely relates to routine clinical practice, it makes that the link between initial presentation and final outcome is not straightforward, given that, at the clinicians discretion, interventions have been done during labour that might have been unnecessary to

prevent adverse outcome. Therefore, any algorithm proposed to act as a backbone for a decision support system (DSS) should take this nonstandardised decision making during the study data collection into account.

4.1.2. Machine learning in a clinical setting

In personalized medicine, the treatment of each specific patient is tailored towards their specific needs, based on the available data and relevant information previously learned from clinical trials and cohorts. This requires the detailed characterisation of the patient and the determination of what can be referred to as their "phenotype" [Cikes et al., 2019]. Many clinical conditions are observed to manifest heterogeneously among different patients, when a thorough data collection is carried out. Additionally, not only will a patient show a particular phenotype at first presentation to the clinician, but over time and with different interventions performed, data will continuously change (reflecting improvements or worsening of their condition).

Machine learning (ML) approaches have been applied in attempts to ease the implementation of personalised medicine, specifically in the fields of diagnosis, classification, prognosis and treatment selection of many conditions. Supervised ML uses algorithms that "learn" from large (accurately) labelled training datasets. However, many of the proposed approaches are difficult to explain with regards to the (clinical) reasoning that is used. On the other hand, unsupervised ML does not aim at providing an answer to the specific learned question (diagnostic label, prognosis, etc.) but instead groups individuals based on their characteristics as described by the available (heterogeneous and rich) input data. Through this grouping, clusters of similar patients can be identified, and common characteristics/"phenotypes" can be described and linked to diagnostics, treatment response, and so forth.

Recently, ML, especially "deep learning", has shown to be very successful at tasks where there is a clear "ground truth" for learning,

such as the identification of objects (e.g. cats, fruits) from pictures and, in medicine, the segmentation of well-defined structures (such as the cardiac cavities from medical images). However, when used for decision-making in clinical practice, ML can be more problematic. A well-known example is that of risk assessment and prognosis in pneumonia, where ML algorithms classify patients with asthma as being of low-risk and not needing interventions, while this finding was based on the fact that, in the learning dataset, asthma patients with suspected pneumonia were treated much more aggressively at presentation, thus effectively lowering their risk when compared to non-asthmatics [Ambrosino et al., 1995]. As ML predictions can rely on or reveal important biases in the population, one needs to interpret results prior to applying it within clinical workflows.

There are two approaches that can ensure that ML does not lead to unwanted results/increased risk for the patient. On the one hand, explicitly intelligible models can be used that allow interpretation and removal of unwanted effects [Caruana et al., 2015] and, on the other hand, a more intuitive approach can be used where patients are compared to one another and the computed similarity is subsequently presented to the clinician and used for prognosis and therapy predictions [Cikes et al., 2019]. In both cases, in order to end up with a tool that can be deployed in a real-world clinical setting, it is crucial that it is not a stand-alone black-box system, but that it integrates well in the framework of decision-making by clinicians/is interpretable as a sequence of clinically meaningful criteria. Figure 4.1 illustrates this (as example in a cardiology setting) and compares ML approaches to what experienced clinicians would do in clinical practice. Clinicians would explore all available data of a given patient and use experience to compare the whole of this data to those of patients they have seen before or were trained to recognise. After positioning the individual with regards to "normality" and typical cases, the course of action is defined based on previous experience regarding treatment results. Given that this is an "eminence-based" subjective approach that only works well for experienced clinicians, many professional organisations

(including the WHO) have provided guidelines that, based on what was observed in large cohorts or clinical trials, make recommendations for patient management. Although these recommendations have proven to standardise medical care in a better way, there are still issues and room for improvement. Most importantly, the current process of formulating guidelines is not guaranteed to make the best use of the original data from trials and studies it is based on. Here, two aspects are relevant: firstly, the fact that studies often skip a thorough analysis of the complex original data (e.g. when images, physiological measurements or lab-analysis, are available) and start off from established derived simplified measures; secondly, clinical practice is often much more complex and varied than clinical trials, where well-selected patients are included to be managed by well-defined protocols of care. In clinical reality, patients often present outside the narrow selection criteria of trials (e.g. regarding co-morbidities, ethnicity, gender, age, lifestyle) or at a different stage of disease/condition; they might have been treated before using different protocols; the acquisition of certain types of data might not be feasible owing to lack of resources, and so forth.

Given the power of contemporary ML for other applications, one can wonder how it can be helpful in providing a monitoring-to-action tool to support the clinician in decision making with real-world patients, or to aid in the extraction/reformulation of "best practice", through learning from full datasets of trials/cohorts or even the combination of different datasets.

As mentioned above, the blind application of ML on large datasets without in-depth knowledge of exactly what it clinically represents, which bias might be embedded, and which approach to decision-making was used, can lead to unwanted results, being potentially dangerous for the patient. A more promising approach is based on data dimensionality reduction, where all complex data is first used in an agnostic way to identify the most relevant features to describe a population, and after which all individuals are ordered according to similarity of these features. In this new ordering, individuals are

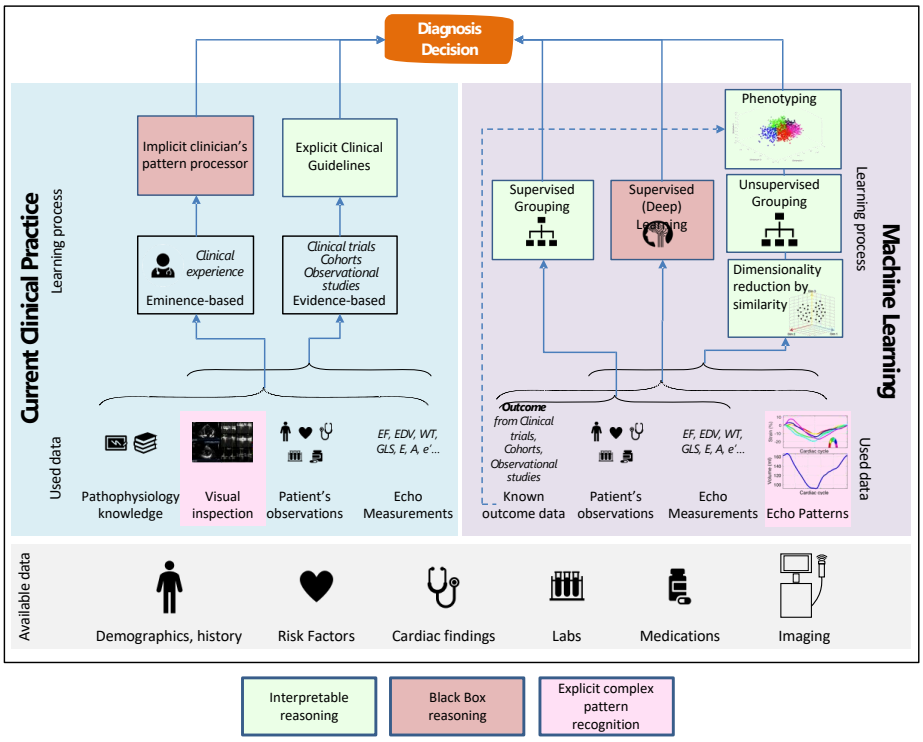


Figure 4.1: Machine learning in clinical decision making (adapted from [Cikes et al., 2019]).

close to each other if they clinically present in a similar way and far from each other otherwise. While this can be used for diagnostic labelling with different gradations of normality-abnormality, it also provides an intuitive approach towards the assessment of therapies and interventions, given that these are aimed to transition an individual towards increased "normality", while taking into account demographics, clinical history, and others.

We have recently shown that an unsupervised approach implementing this idea, based on multiple kernel learning (MKL), can provide useful insight in (large) complex patient populations [Sanchez-

Martinez et al., 2018, Nogueira et al., 2020a]. Additionally, this approach can be used to study temporally dynamic phenomena where the condition of a patient changes over a (short or long) period of time as a result of an "intervention" [Nogueira et al., 2020a].

In this report, we describe BCN-SELMA, our implementation of a *Simplified, Effective, Labour Monitoring-to-Action* tool based on this idea of similarity-based dimensionality reduction.

4.2. Methods

4.2.1. Cohort

The proposed approach, as well as its validation, is based on the SELMA project [Souza et al., 2015], a multicentre (9 major hospitals in Nigeria and 4 in Uganda) study conducted by the WHO and aimed at providing insight and new tools towards the reduction of labour-associated maternal, fetal and neonatal mortality and morbidity. The study included the collection of a prospective cohort of 9995 women, with data being collected at admission to the centre, throughout labour and after birth [Souza et al., 2015]. Intrapartum care was provided based on standard clinical guidelines for good obstetric care practices [National Institute for Health and Clinical Excellence., 2007, World Health Organization., 2014] and by skilled professionals with free access to labour intervention (caesarean section (CS), augmentation, assisted vaginal delivery, etc.) resources. An aggregated bad outcome (BO) was defined as the composite of stillbirth, intra-hospital early neonatal death, neonatal use of anticonvulsants, neonatal cardio-pulmonary resuscitation, Apgar score below 6 at 5 minutes, uterine rupture, maternal death or organ dysfunction preceded by dystocia.

Table 4.1 summarises the outcome and main interventions performed in this cohort.

Table 4.1: Summary of outcome and main interventions in SELMA study. BO = Bad Outcome; CS = Caesarean Section.

		Number of patients (out of total 9995) (%)
BO	Total	2.23
	Fetal/Neonatal	2.01
	Maternal	0.26
CS		13.31
Augmentation		35.08

4.2.2. ML Approach

4.2.2.1. Overall approach

As discussed in 4.1.1, given that the data show a severe imbalance in BO, a large diversity in admission variables (e.g. women enter the study in different stages of labour), and that the temporal monitoring is totally dyssynchronous, it is clear that there is no "straightforward" ML-based tool to predict the adverse outcome from the admission data. Additionally, given that the interventions are performed to prevent adverse outcome, but without clear decision criteria, there is no unique and well-defined path for each patient from admission to final outcome.

As discussed in 4.1.2 , it is important that the chosen approach is intrinsically integrated in the traditional workflow to manage labour. The latter is illustrated in Figure 4.2.

When a woman presents to the facility to give birth, admission information is obtained, including maternal characteristics and a first assessment of the stage of labour and the fetus. Next, an iterative process is started where mother and fetus are assessed. In each iteration, either the timing of the next reassessment is chosen or, if determined appropriate, an intervention is performed to adjust labour progress. This process continues until birth (and potentially further on until discharge from the facility). In BCN-SELMA, we have chosen

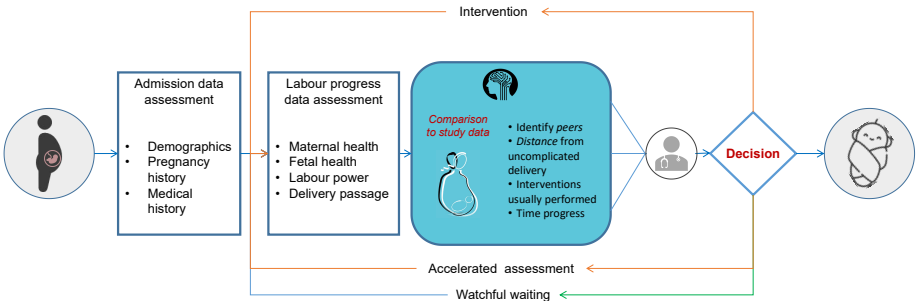


Figure 4.2: The overall approach for ML-based management of labour.

the following approach (Figure 4.3):

- build a low-dimensional space, positioning individuals based on the similarity in their admission data using the MKL algorithm. In practice, a partition of the SELMA dataset was used to calculate the admission-based low-dimensional space.
- the dynamic data is projected onto the learned space, and temporal trajectories of each individual are quantified and linked to performed interventions and final outcome.
- when a new mother presents, she is first positioned in this space based on admission data.
- during labour, her temporal trajectory is continuously updated and compared to those from known peers to determine “deviation from normality”, as well as to calculate the chance of intervention based on what was done in the SELMA training set.

Herein, we briefly summarise the processes of building the MKL space with the baseline SELMA data and dynamic data analysis. For a more detailed description, we refer the reader to [Nogueira et al., 2020b].

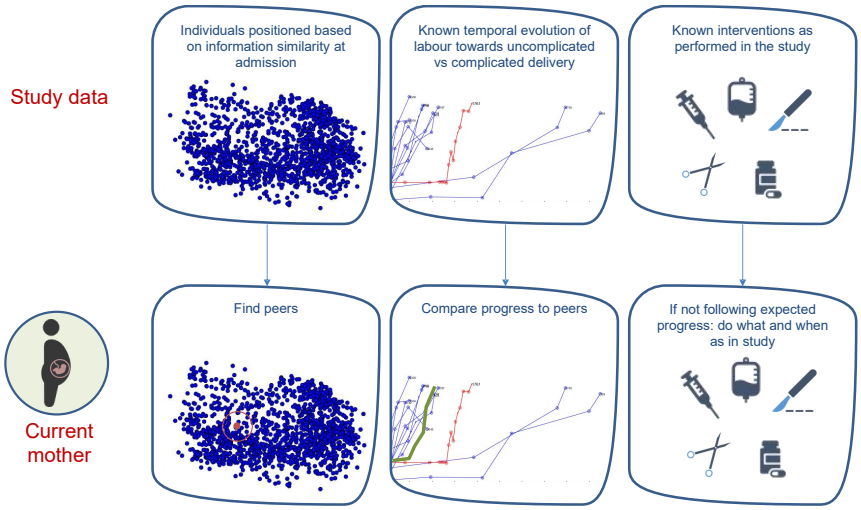


Figure 4.3: Illustration of the algorithm on behind BCN-SELMA.

4.2.2.2. Building the MKL space with baseline data

The building of the MKL space with baseline data involved three main steps: (1) data preprocessing and partitioning, (2) preparation of MKL algorithm for high performance computing, and (3) actual learning of the space.

I. Data preprocessing and partition. From the numerous variables available from the study, the most relevant features for an initial positioning of patients with regards to each other were identified, in cooperation with clinical experts. A selection of 52 features was used to characterize each patient, at any time point. Of those 52 features, 33 corresponded to parameters that were acquired only at admission, and are mostly related with patient demographics, medical background and other characteristics that remain unchanged during the course of labour (see Table 4.5 of Appendix 4.A). The remaining 19 features were followed-up during the process of labour, in non-

standardised intervals (see Table 4.6 of Appendix 4.A). Some of the features were directly used in their original SELMA dataset form, whereas others result from some type of processing/combination of several others. Those that do not appear in their original form have their names emphasized in bold.

As previously referred, the original dataset consists of 9995 patients. Often, values were missing from different features/follow-ups of each patient. Missing data was mainly dealt with by previous (follow-up) value propagation. Cases with important missing admission data were discarded from the analysis. A total of 9446 patients remained. They were divided into training (75%) and testing (25%) partitions. The incidence of the main labour-associated interventions and adverse outcome was verified to be identical in the two partitions.

The MKL projection model was learnt from the admission/first assessment data of the training portion (7085 patients), and the testing portion (2361 patients) was used to simulate the occurrence of new cases.

II. Parallelisation and execution using high performance computing. In MKL, as in other kernel-based methods, scalability is an issue: as data size increases, memory and time requirements quickly become intractable. With the current available implementations [Sanchez-Martinez et al.,], running the algorithm with a dataset of this size would imply several days for a single iteration. To improve scalability, parallelisation strategies addressing the most computationally expensive steps were developed, combining message passing interface (MPI) [Gropp et al., 1996] and open multi-processing (OpenMP) [Dagum and Menon, 1998].

The computations were performed on the NORD III supercomputer of the Barcelona Supercomputing Center [Barcelona Supercomputing Center,], using up to 17 computing nodes (IBM dx360 M4), with 16 cores each (2x Intel SandyBridge-EP E5–2670 2.6GHz cache 20MB 8-core). Each node is equipped with a total RAM of 128GB (8x 16G DDR3–1600 DIMMs (8GB/core)) and 500GB of disk storage

(7200 rpm SATA II HDD).

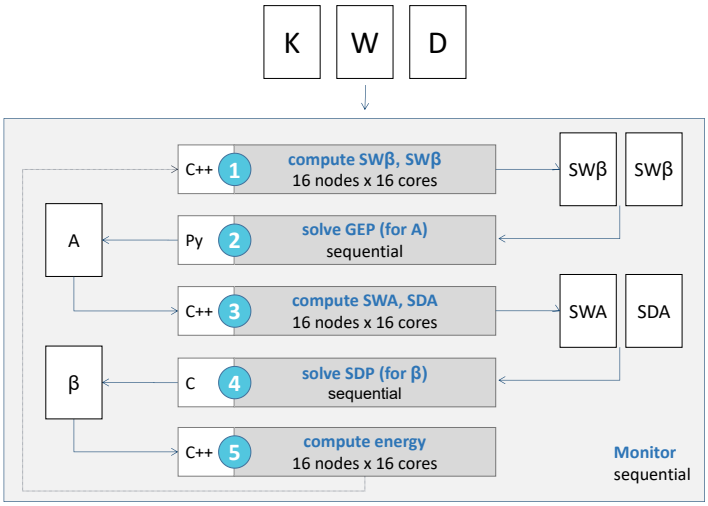


Figure 4.4: High-level illustration of the MKL implementation. One full iteration corresponds to a two-step optimization, the first step consisting of a generalized eigenvalue problem (GEP) – to solve for projection matrix A – and the second step consisting of a semidefinite programming problem (SDP) – to solve for feature weight vector β. We divided it in 5 blocks/standalone jobs: 1 - computing the matrix pair of the GEP; 2 - solving the GEP; 3 - computing the input matrices of the SDP; 4 - solving the SDP; and 5 - computing the energy. For matrix notation and detailed job descriptions see [Nogueira et al., 2020a, Lin YY, 2011].

The 5 main steps that make up one iteration of MKL were implemented as standalone jobs with specific inputs and outputs (see high-level illustration in Figure 4.4). A monitor application is responsible for submitting job $k+1$ only after job k is successfully completed, assessing convergence at the end of job 5, and either stopping execution or resuming with job 1 for the next iteration. Data exchange between consecutive jobs is handled via storing and reading of binary files. The monitor application was developed in C++, as well as jobs 1, 3 and 5. The C++ *Eigen* library [Guennebaud et al., 2010] was

chosen for handling matrix operations. As for jobs 2 and 4, open source solvers are currently being used – Python *scipy*'s *eig* [Virtanen et al., 2020] and the C library *CSDP* [Borchers, 1999]. These 2 jobs execute relatively fast, so we focused our efforts in the parallelisation of jobs 1, 3 and 5, the real bottlenecks of the algorithm. A hybrid MPI-OpenMP strategy was utilized. All 3 jobs perform a set of $\approx \frac{N \times N}{2}$ (with N = number of cases) independent operations whose results are summed, the result of interest being the global sum. We used the MPI protocol to distribute the computations across 16 nodes: the master node reads and broadcasts the necessary input data to the other nodes; then, each of the nodes computes a partial sum. After all nodes return their partial sums, the master performs the global sum. Within each node, OpenMP threads are used to distribute the partial sums by the available cores.

III. Building the MKL space. Using this implementation, the baseline variables were finally converted into coordinates in the low-dimensional representation. MKL represents all features in a unified manner (through kernel/similarity matrices), and all further operations are performed on the kernelized (instead of raw) data. A global similarity matrix, which results of a combination of all features' individual similarity matrices, dictates how overall (dis)similar each two patients are, which ultimately dictates their distancing in the low-dimensional space. The projection model consists of a projection matrix A and a feature weight vector β, and can be straightforwardly used to project new data (i.e. "unseen" in the learning). The approach is illustrated in Figure 4.5 and described in detail in [Lin YY, 2011, Sanchez-Martinez et al., 2017].

4.2.2.3. Dynamic data analysis

After learning the coordinates of the baseline data of the "training" individuals in the training set, the MKL projection model can be used to project their follow-up data. After this step, each individual

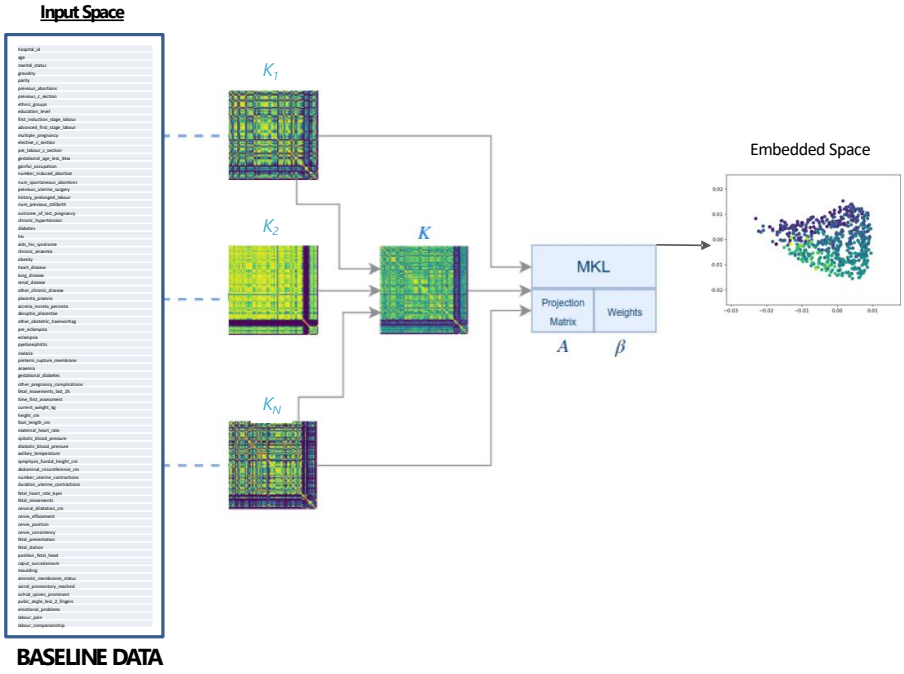


Figure 4.5: From the relevant baseline variables, similarity matrices are computed quantifying pairwise similarity of individuals for each feature. Using MKL, a low dimensional space is constructed positioning each individual with regards to each other based on similarity measures of the baseline variables.

is associated with low-dimensional trajectory (Figure 4.6a). Likewise, any new data, from individuals that were not used to train the model, can also be projected onto the MKL space.

The management of new individuals starts with their positioning in the low-dimensional space based on admission variables. At each follow-up, their positions are updated. Each time a patient's position changes in the low-dimensional space, the "training" patients whose projections lie in a close neighbourhood (peers) are retrieved and used to estimate the "ideal" future trajectory (the average trajectory

among peers with complication- and intervention-free labours, see Figure 4.6b). This way, individual trajectories can be compared and interpreted with regards to deviation from "normality". Additionally, information on the interventions that were performed among peers can be used to estimate a chance/risk of intervention.

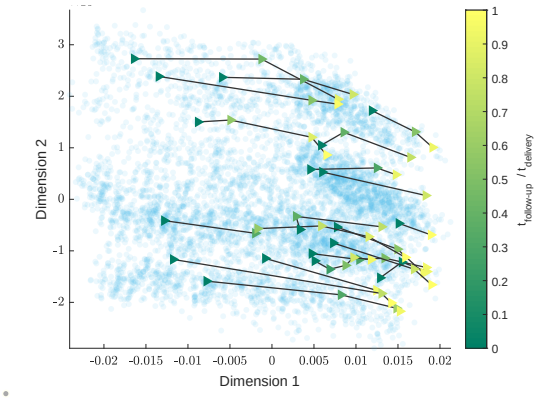
Therefore, at each point in time, we can identify how labour is progressing with respect to "normality" and what type of interventions are recommended to prevent adverse outcome according to SELMA practice. The methodology for dynamic data analysis is described in more detail in [Nogueira et al., 2020b].

4.2.3. Decision Support Tool

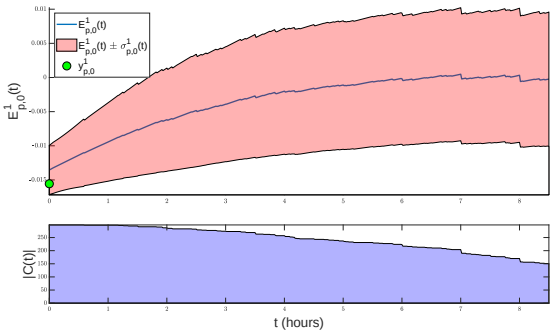
4.2.3.1. Overall approach

The ultimate objective of this work is to materialize the previous methodological pipeline into a DSS that is deployable in a clinical environment. Once a database of low-dimensional trajectories of "training" patients is available, in order to handle the management of a new patient, the DSS would require 6 main components, each associated with one of the following tasks:

1. To capture the admission data, the follow-up information while labour progresses, intervention and outcome information.
2. To position (=project) the new patient within the low-dimensional space, learned from the training data.
3. To calculate the expected trajectory for the new patient during labour, as well as to compare the trajectory of the new patient to the ideal one.
4. To provide an estimate of the chance of adverse outcome.
5. To provide an estimate of the need for a certain intervention, that could reverse deviation from the predicted optimal path, at a certain time point in the future during labour progress.



(a)



(b)

Figure 4.6: Trajectories of individuals in the low dimensional space during labour. (a) Individual trajectories from admission until delivery. (b) Estimation of the expected path for an individual based on peers with an uncomplicated delivery (top) and the number of uncomplicated peers used for this estimate as a function of time (bottom).

6. To dynamically update the position of the new patient within the low-dimensional space when new measurements become available during labour.

4.2.3.2. Implementation

Each of these components have been implemented in a web- and cloud-based prototype of the BCN-SELMA DSS.

Component 1. Data entry is performed over a spreadsheet-like form (Figure 4.7– right side). This sheet has three tabs: one for entering the admission variables, one for entering the dynamic variables during labour progress, and a third one for entering intervention and outcome information. When data are entered, they are stored in the database of the system. In order to speed up data entry, if there are variables that are very likely to show specific initial values, they are pre-filled with such values as soon as data is entered in one of the obligatory fields. Suggested information is displayed in orange, whereas effectively entered data is shown in green. Additionally, there is a range check for the variables based on the expected values from the SELMA study, with out-of-range data being shown in red.

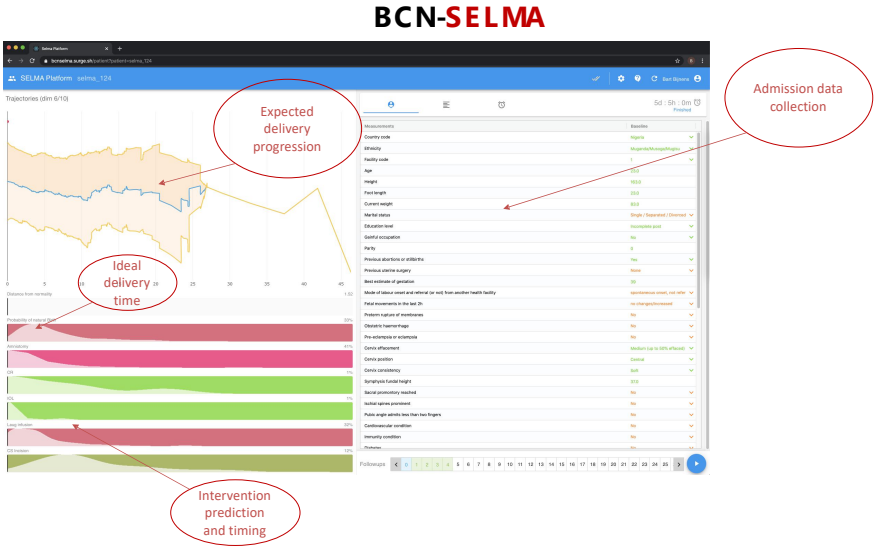


Figure 4.7: Initial view for admission data entry and a first estimate of trajectory over time, as well as interventions to be performed.

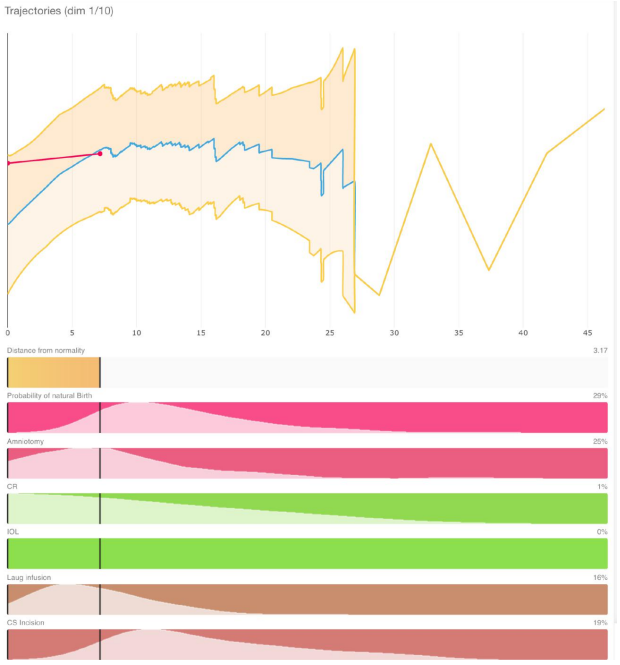


Figure 4.8: Detail of the prediction panels of the web interface.

Component 2. As soon as the required admission information has been entered, the system is triggered to project this information into the learned MKL space. This is performed by the cloud-based computation engine (discussed ahead). In the initial prototype, this is based on MATLAB code that is executed in a Docker environment.

Component 3. Once the projection of the patient is known, the computation engine will calculate their “ideal” labour progress trajectory , again by executing MATLAB code inside a Docker container. The peers from the training set in a certain (multidimensional) neighbourhood of the position of the new patient (see [Nogueira et al., 2020b] for details) are determined and split into uncomplicated births versus those with interventions or adverse outcomes. Then, the average trajectory and corresponding standard-deviation, for the uncomplicated peers, are calculated. The results are visualised in the

left (upper) panel of the web-interface, as illustrated in Figure 4.7. Given that this is a trajectory in a high-dimensional space, the dimension that is displayed in the interface (versus time) is the one for which the patient currently deviates the most.

As soon as the ideal trajectory is known, the prognosis estimation for time of birth, adverse outcome, and all possible interventions, can be performed (**components 4 and 5**). These estimates are visualised in the prediction field (lower left panel) of the web-interface (Figure 4.7, detailed in Figure 4.8). All backend calculations of **components 4 and 5** are also executed within a "MATLAB Docker".

Component 4. The top bar shows the distance of the current patient to the "ideal" trajectory at each time-point. The second bar plots the distribution of the time of delivery of the uncomplicated births. Additionally, the chance of a fully uncomplicated birth is calculated as the percentage of peers that had one, and it is used to set the bar's background colour (green for low chance and red for high chance). The actual value is printed to the side of the bar. Together, these two bars (distance from the normal trajectory and chance of an uncomplicated delivery) provide an estimate of risk of adverse outcome.

Component 5. The next bars of the web-interface are used to visualise the chance of a certain intervention (in this prototype: amniotomy, cervical ripening, induction of labour, labour augmentation, and caesarean section) and most likely timings. The chance/risk of performing the intervention is estimated as the incidence rate among peers, sets the background colour of the bar and is printed on the side. Again, 0% chance maps to green and 100% chance maps to red. The temporal distribution of the intervention among peers is plotted in a lighter colour (Figure 4.8).

Component 6. Dynamic data are similarly entered in the appropriate tab (Figure 4.9). Each time a new value is entered (after a certain time period of the previous one), a new time-point (=column in the spreadsheet) is created and timestamped. Here, for most variables, the value of the previous capture is copied to speed up

data entry and ensure completeness. Once the required variables are available, a "MATLAB Docker" recalculates the position of the patient in the MKL space.

As soon as the new position is available, all elements of the left panels (trajectory plot and intervention/outcome bars) are updated. Figures 4.7-4.11 illustrate this dynamic process for a patient that finally underwent a CS.

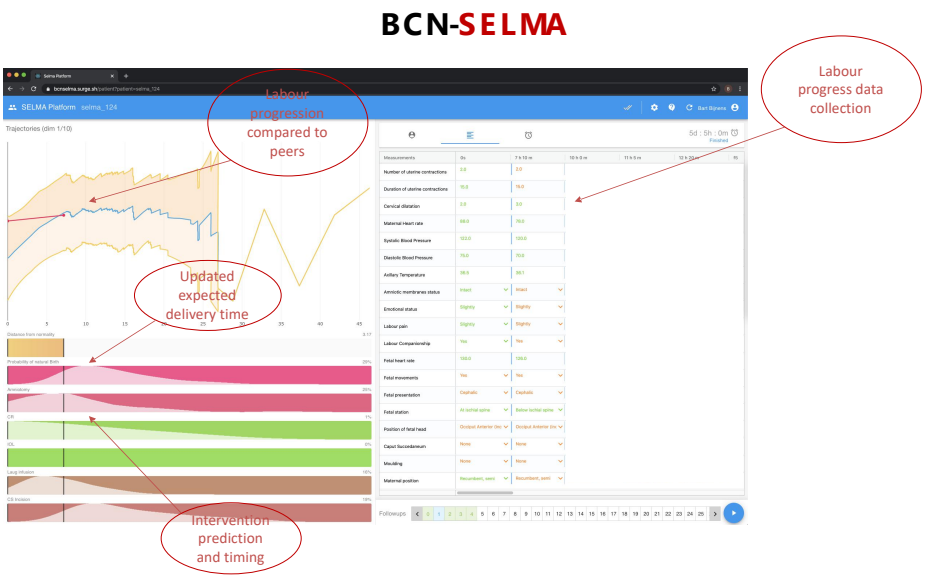


Figure 4.9: Once a new set of dynamic data becomes available, an update of trajectory and estimations is shown.

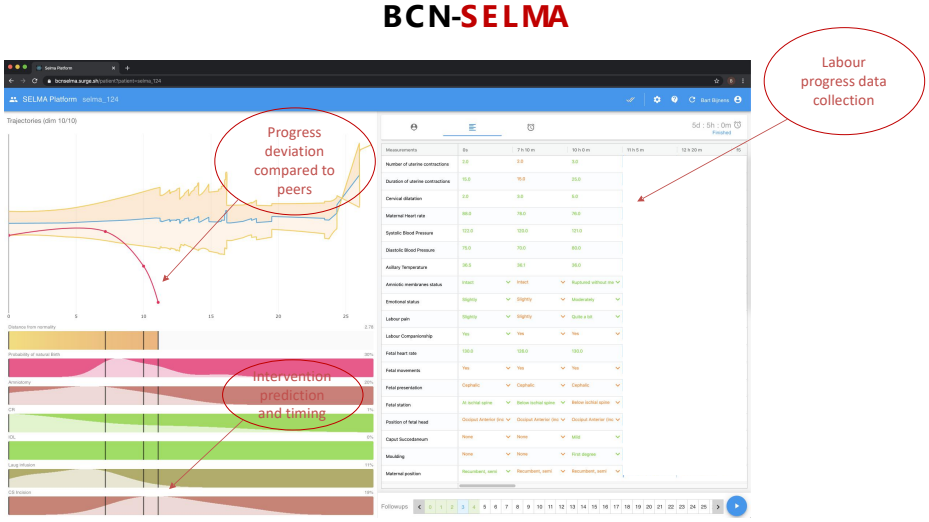


Figure 4.10: In this example, at this stage of labour, the distance from the normal trajectory is large, most peers would have given birth already and the system predicts CS with the highest probability.

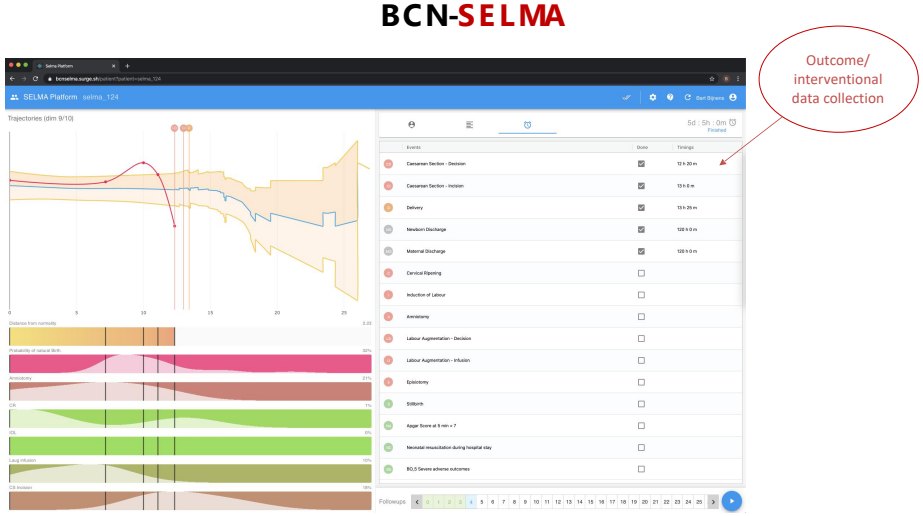


Figure 4.11: In this example, a CS was effectively performed, and all outcome and intervention data is recorded in the appropriate tab.

4.2.3.3. Infrastructure

As previously mentioned, the prototype of the BCN-SELMA DSS is implemented as a web- and cloud-based platform. The current implementation is made of the components and interactions illustrated in Figure 4.12. The platform is based on virtual servers on a private Kubernetes cloud, where the different independent components (Dockers) are managed. The main components are:

- The computation engine: the part of the platform in charge of executing the patient specific projections and predictions.
- Databases: where (anonymised) patient data, as well as the learned MKL space and parameters relevant for the executions, are stored.
- Application programming interfaces (APIs): the communication interfaces between the web client (user) and the platform. These components allow the user to save and get data from databases and execute the predictions.
- Third parties: the current prototype uses the Google authentication approach to provide safe access to the platform and identify the different users (allowing to store all individual transaction/access information for auditing the system).
- Proxy: the technology in charge to expose the APIs to the internet using a DNS or URL.

In order to guarantee optimal data security, all the data in the platform are pseudo-anonymised, with real ID's and vulnerable information remaining in the original source, where the care is provided, and not in the platform. Once the user accesses the platform, all the communication between the computer and the cloud platform is done using an encrypted HTTPS channel. The platform uses two authentication strategies before accessing any data, Google Auth and JWT authentication.

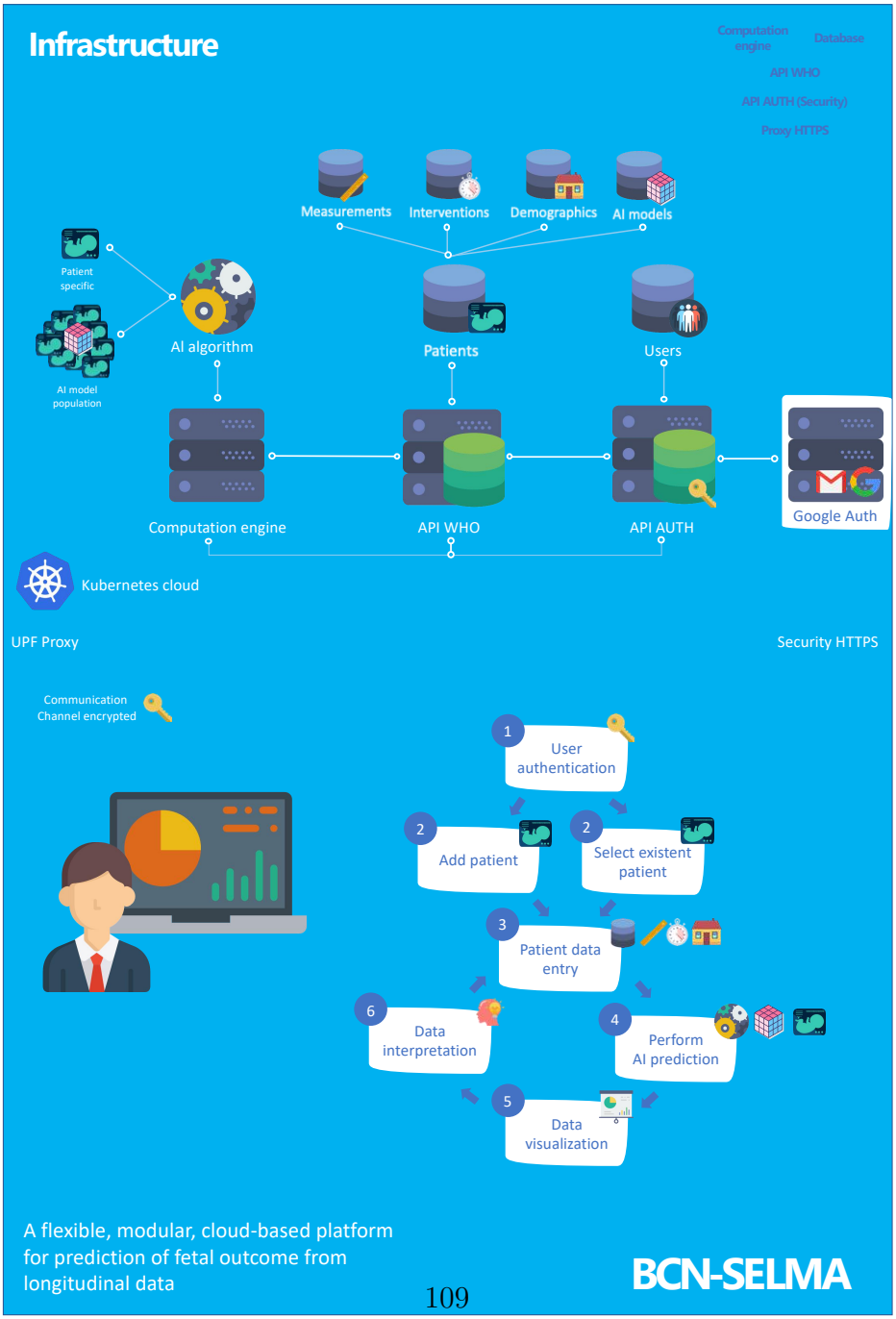


Figure 4.12: The infrastructure on which the prototype of BCN-SELMA is implemented.

The platform can be accessed through a standard web browser and is implemented in JavaScript (for the user-interface and interaction, as well as the backbone of the platform), whereas the projection and prediction parts are based on dockerised MATLAB code.

4.2.4. Evaluation

As described above, the SELMA dataset was subdivided in training and testing sets. The training set is used to calculate the MKL space, the estimated trajectories and interventions to be performed. For each of the test individuals, the above DSS is executed during the whole labour progress (illustration in Figure 4.13). The ability of the system to predict adverse outcome, as well as CS (as example of intervention) is subsequently evaluated.

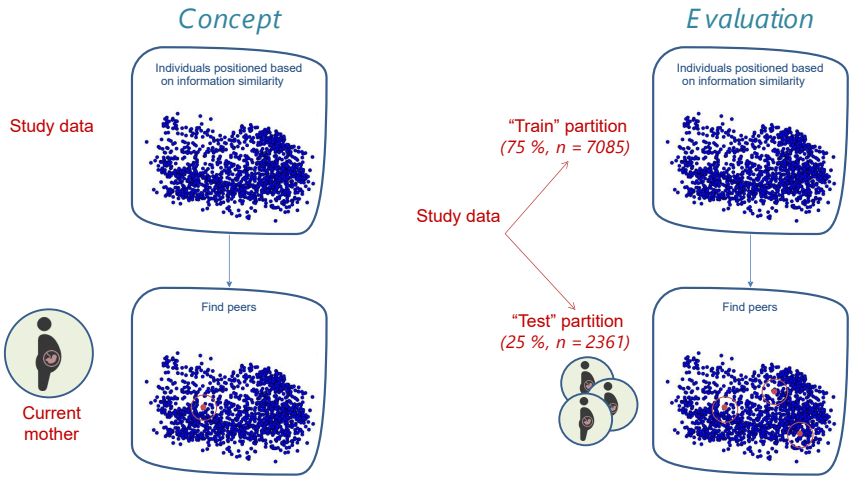


Figure 4.13: Evaluation of the performance of the BCN-SELMA prototype.

The different parameters used for prediction were:

- v_π: the chance that a certain event/intervention happened amongst peers of the individual to evaluate.

- $v_{\pi z}$: the chance of an event/intervention combined with maximal distance from normality of the temporal trajectory.
- $v_{\pi zt}$: the chance of an event/intervention, combined with the maximal distance from normality as well as the time elapsed since admission.

Decision thresholds were learned for each parameter, using a cross-validation scheme on the training set (illustration in Figure 4.14). A more detailed description of the evaluation pipeline can be found in [Nogueira et al., 2020b].

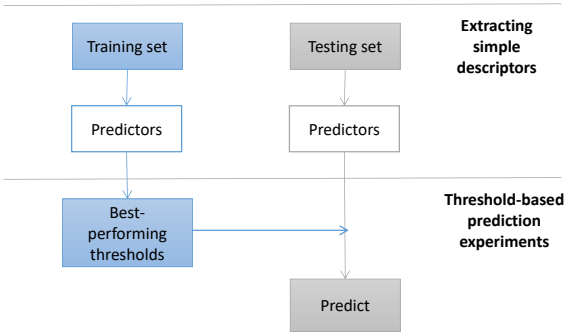


Figure 4.14: The approach to determine the prediction thresholds from the training data.

4.3. Results

MKL calculation. As can be seen in Table 4.2, we managed to achieve an average full-iteration time of under 1 day and a half. The algorithm ran for 20 iterations, taking a little over 28 days to completion.

Decision support system. The performance of the DSS is documented in more detail in [Nogueira et al., 2020b]. Tables 4.3 and 4.4, respectively, show the performance to predict CS and adverse outcome using different decision approaches and as compared to the partograph.

Table 4.2: Average execution time (in hours) for each job for one MKL iteration, one full iteration and total time.

JOB 1	**JOB 2**	**JOB 3**	**JOB 4**	**JOB 5**
3.25 ± 0.09	0.98 ± 0.06	28.4 ± 0.04	0.0012 ± 0.0007	1.1644 ± 0.1106

Full iteration	**20 Iterations**
33.8 ± 0.14	676.3 (≈ 28 days)

Table 4.3: The performance of the BCN-SELMA prototype to predict caesarean section. n = sample size; n_{CS} = number of positive cases; Th = threshold/cut-off; SE = sensitivity; SP = specificity; PPV = positive predictive value; NPV = negative predictive value; AUC = area under the receiver operating characteristic; p-value = fraction of random permutation tests for which $AUC \geq AUC_{observed}$ (total of 10000).

Train ($n = 6349$; $n_{CS} = 817$)						
	Th	**SE**	**SP**	**PPV**	**NPV**	**AUC (p-value)**
Alert line	-	0.540	0.728	0.227	0.915	-
Action line	-	0.290	0.889	0.278	0.894	-
v_{π}^{CS}	0.221	0.699	0.700	0.256	0.940	0.763 (< 0.0001)
$\mathrm{v}_{\pi z}^{CS}$	0.422	0.683	0.684	0.242	0.936	0.746 (< 0.0001)
$\mathrm{v}_{\pi zt}^{CS}$	2.038	0.706	0.707	0.263	0.942	0.767 (< 0.0001)
Test ($n = 2121$; $n_{CS} = 279$)						
	Th	**SE**	**SP**	**PPV**	**NPV**	**AUC (p-value)**
Alert line	-	0.548	0.731	0.236	0.914	-
Action line	-	0.290	0.891	0.288	0.892	-
v_{π}^{CS}	0.221	0.674	0.696	0.251	0.934	-
$\mathrm{v}_{\pi z}^{CS}$	0.422	0.659	0.712	0.258	0.932	-
$\mathrm{v}_{\pi zt}^{CS}$	2.038	0.703	0.712	0.270	0.941	-

Table 4.4: The performance of the BCN-SELMA prototype to predict adverse outcome. n = sample size; n_{BO} = number of positive cases; Th = threshold/cut-off; SE = sensitivity; SP = specificity; PPV = positive predictive value; NPV = negative predictive value; AUC = area under the receiver operating characteristic; p-value = fraction of random permutation tests for which $AUC \geq AUC_{observed}$ (total of 10000).

Train ($n = 6349$; $n_{BO} = 155$)						
	Th	**SE**	**SP**	**PPV**	**NPV**	**AUC (p-value)**
Alert line	-	0.419	0.697	0.033	0.980	-
Action line	-	0.174	0.867	0.032	0.977	-
v_{π}^{BO}	0.036	0.594	0.594	0.035	0.983	0.612 (< 0.0001)
$v_{\pi z}^{BO}$	0.069	0.561	0.567	0.031	0.981	0.581 (0.0008)
$v_{\pi zt}^{BO}$	0.283	0.568	0.573	0.032	0.981	0.595 (< 0.0001)
Test ($n = 2121$; $n_{BO} = 44$)						
	Th	**SE**	**SP**	**PPV**	**NPV**	**AUC (p-value)**
Alert line	-	0.455	0.698	0.031	0.984	-
Action line	-	0.205	0.869	0.032	0.981	-
v_{π}^{BO}	0.036	0.523	0.635	0.029	0.984	-
$v_{\pi z}^{BO}$	0.069	0.500	0.605	0.026	0.983	-
$v_{\pi zt}^{BO}$	0.283	0.568	0.557	0.026	0.984	-

4.4. Conclusion

In this project we have created a prototype of a *Simplified, Effective, Labour Monitoring-to-Action tool, based on Interpretable Machine Learning*, centered on the idea of dynamically identifying peers of the individual to monitor and using the available peer information regarding evolution, interventions and outcome to provide a personalised monitoring and dynamic estimate of need for intervention/risk of adverse outcome.

The prototype was implemented in a web- and cloud-based environment with an intuitive user-interface that has a look-and-feel similar to a personalised partograph that can be dynamically used to manage a new delivery.

We have performed a classical performance evaluation to predict adverse outcome as well as CS and have shown that the proposed

approach leads to a higher, as well as a more balanced, sensitivity and specificity as compared to the partograph.

The proposed BCN-SELMA prototype can thus be seen as the basis for as a personalised, evidence-based, labour monitoring tool, that has an intuitive user-interface, enabling fast visual assessment of labour progress. The approach allows customization based on local data, incorporating local guidelines and practice. The proposed implementation as a web- and cloud-based platform allows scalability and flexible deployment as well as update of the underlying algorithms.

Appendices

4.A. Static and Dynamic Features

Table 4.5: Admission-only / static features.

	NAME	NOTES
1	Country code	Country code: Uganda/Nigeria (1/0)
2	**Ethnicity**	Ethnicity: [NIGERIA] 1 - Ibo; 2 - Yoruba; 3 - Hausa; 4 - Fulani; 5 - TIV; 6 - Kanuri; 7 - Other Nigerian; 8 - Non Nigerian; [UGANDA] 9 - Muganda/Musoga/Mugisu; 10 - Munyakore/Mukiga/Munyoro/Mutoro; 11 - Acholi/Langi/Alur; 12 - Iteso/Karamojong; 13 - Lugbara/Madi; 14 - Other Ugandan; 15 - Non-Ugandan
3	Facility code	1-13
4	Age	years
5	Height	cm
6	Foot length	cm
7	Current weight	kg
8	Marital status	Marital status: 0 - Single / Separated / Divorced / Widowed; 1 - Married / Cohabitating

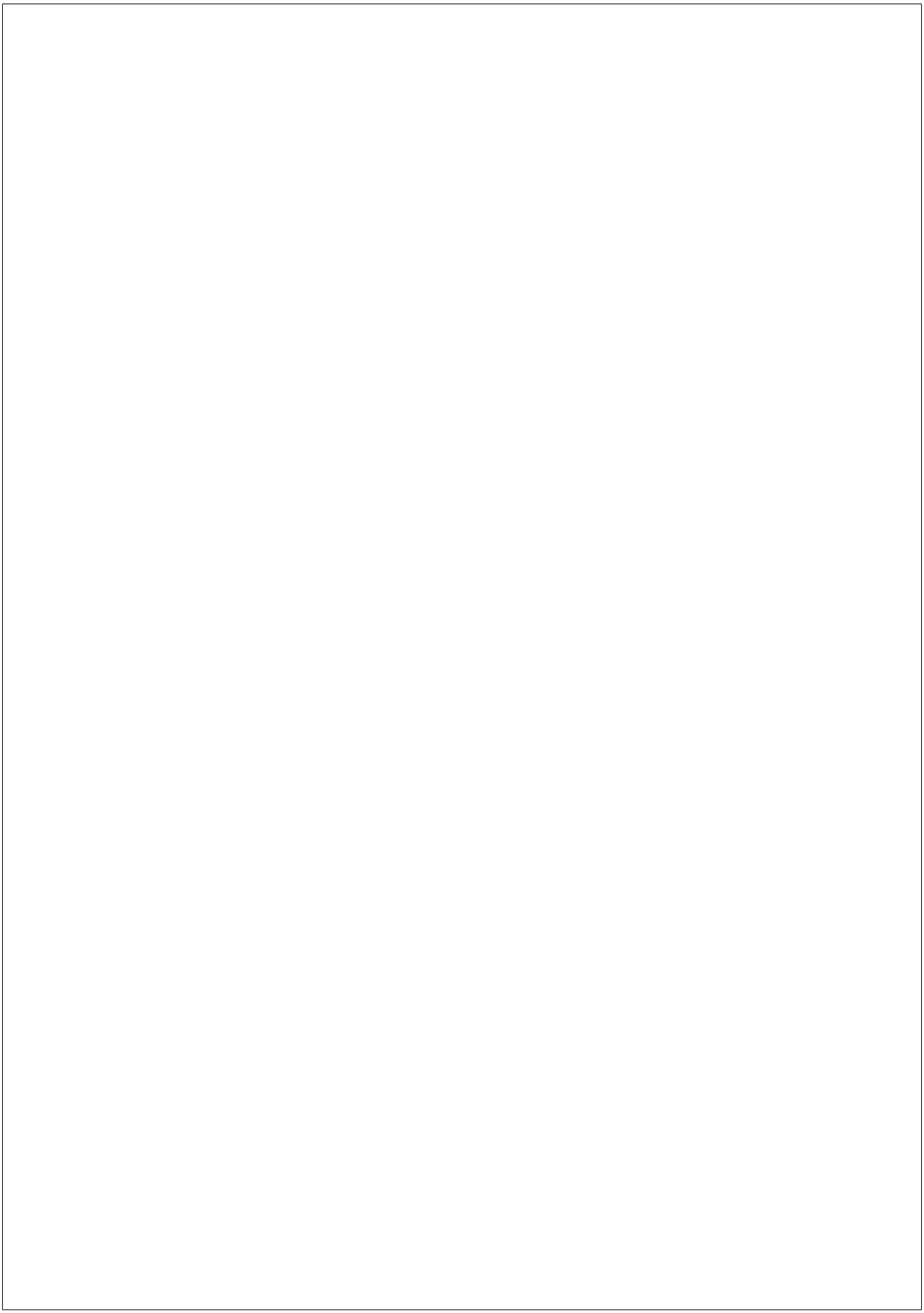
9	**Education level**	Education level: 0 - No education; 1 - Other (e.g. Quranic / Nomadic education only; 2 - Pre-primary education; 3 - Incomplete primary education; 4 - Complete primary education; 5 - Incomplete secondary education; 6 - Complete secondary education; 7 - Incomplete post-secondary/tertiary education; 8 - Complete post-secondary/tertiary education)
10	Gainful occupation	Gainful occupation: 0 - No; 1 - Yes
11	Parity	Number of previous births
12	**Previous abortions or stillbirths**	Previous abortions or stillbirths: 0 - No; 1 - Yes
13	**Previous uterine surgery**	Previous uterine surgery (includes previous c-sections or other uterine surgeries): 0 - None; 1 - One; 2 - More than one
14	Best estimate of gestation	weeks
15	**Mode of labour onset and referral (or not) from another health facility**	Mode of labour onset and referral (or not) from another health facility: 0 - spontaneous onset, not referred from another facility; 1 - induced, not referred; 2 - spontaneous, referred; 3 - induced, referred
16	**Fetal movements in the last 2h**	Fetal movements in the last 2h: 0 - reduced or absent; 1 - no changes/increased
17	Preterm rupture of membranes	Preterm rupture of membranes: 0 - No; 1 - Yes
18	**Obstetric haemorrhage**	Placenta praevia, accreata increta percreta, placentae abruption or other obstetric haemorrhage: 0 - No; 1 - Yes
19	**Pre-eclampsia or eclampsia**	Pre-eclampsia or eclampsia: 0 - No; 1 - Yes

20	Cervix effacement	Cervix effacement: 0 - Thick (less than 30% effaced); 1 - Medium (up to 50% effaced); 2 - Thin (up to 80% effaced); 3 - Very thin / paper-thin (more than 80% effaced)
21	Cervix position	Cervix position: 0 - Anterior; 1 - Central; 2 - Posterior
22	Cervix consistency	Cervix consistency: 0 - Soft; 1 - Medium; 2 - Firm
23	Symphysis fundal height	cm
24	Sacral promontory reached	Sacral promontory reached: 0 - No; 1 - Yes; 2 - Not assessed
25	Ischial spines prominent	Ischial spines prominent: 0 - No; 1 - Yes; 2 - Not assessed
26	Pubic angle admits less than two fingers	Pubic angle admits less than two fingers: 0 - No; 1 - Yes; 2 - Not assessed
27	**Cardiovascular condition**	Chronic hypertension, heart_disease, obesity, or chronic_anaemia: 0 - No; 1 - Yes
28	**Immunity condition**	HIV or AIDS: 0 - No; 1 - Yes
29	**Diabetes**	Diabetes or gestational diabetes: 0 - No; 1 - Yes
30	**Renal condition**	Pyelonephritis or renal disease: 0 - No; 1 - Yes
31	Lung disease	Lung disease: 0 - No; 1 - Yes
32	Anaemia	Anaemia: 0 - No; 1 - Yes
33	**Other condition**	Other chronic disease, other pregnancy complications, malaria: 0 - No; 1 - Yes

Table 4.6: Follow-up / dynamic features.

	NAME	NOTES
1	Contraction ON time	Duration of uterine contractions (seconds)
2	**Contraction OFF time**	Time between contractions (seconds)
3	Cervical dilatation	cm
4	Maternal Heart Rate	bpm
5	Systolic Blood Pressure	mmHg
6	Diastolic Blood Pressure	mmHg
7	Axillary Temperature	°C
8	Amniotic membranes status	Amniotic membranes status: 0 - Intact; 1 - Ruptured without meconium; 2 - Ruptured with stale meconium; 3 - Ruptured with fresh meconium
9	Emotional status	Since the last assessment, how much the woman has been bothered by emotional problems such as fear, anxiety, depression, irritability, or sadness? 0 - Not at all; 1 - Slightly; 2 - Moderately; 3 - Quite a bit; 4 - Extremely
10	Labour pain	Since the last assessment, how much the woman has been bothered by labour pain? 0 - Not at all; 1 - Slightly; 2 - Moderately; 3 - Quite a bit; 4 - Extremely
11	Labour Companionship	Labour Companionship: 0 - No; 1 - Yes
12	Fetal Heart Rate	bpm
13	Fetal movements	Fetal movements observed/felt: 0 - No; 1 - Yes

14 Fetal presentation	Fetal presentation: 0 - Cephalic; 1 - Breech; 2 - Transverse lie / compound / other
15 Fetal station	Fetal station: 0 - Above ischial spine; 1 - At ischial spine; 2 - Below ischial spine
16 Position of fetal head	Position of fetal head: 0 - Occiput Anterior (includes right and left); 1 - Occiput transverse; 2 - Occiput posterior; 3 - Other
17 Caput Succedaneum	Caput Succedaneum: 0 - None; 1 - Mild; 2 - Moderate; 3 - Severe
18 Moulding	Moulding: 0 - None; 1 - First degree; 2 - Second degree; 3 - Third degree
19 Maternal position	Predominant maternal position between assessments: 0 - Upright, sitting, standing, walking, kneeing, squatting, all-4; 1 - Recumbent, semi-recumbent, lateral, supine



Chapter 5

CONCLUSION

In this thesis, we have developed tools for the analysis of "real-world" longitudinal clinical data while addressing concrete clinical problems. Herein, we summarise the main contributions, limitations and potential future directions, in application-specific and global perspectives.

5.1. Application I: Nonstandardized stress echocardiography

Despite the theoretical benefits of nonstandardized protocols such as the handgrip test in terms of "scalability", a lack of suited analysis tools currently disincentivizes the exploration of their potential. In this application, we illustrated how unsupervised multiview dimensionality reduction (in particular, unsupervised multiple kernel learning (MKL)) could be used towards the analysis of nonstandardized stress echocardiography, while privileging clinical interpretability. The proposed approach allowed to identify normal and abnormal trajectories in response to stress and to relate them with specific changes in the original clinical features. A noteworthy result was the similarity-based positioning of (supposedly) healthy controls and diseased patients in

a spectrum of (ab)normality, rather than in well-separated clusters. This result illustrates why learning based on hard diagnostic labellings might sometimes not be the best approach. In sum, the main contributions of this work are three-fold: (1) the clinical insight regarding the specific data/populations at hand (the characterization of healthy and pathological response patterns), (2) the validation of the proposed methodology as a potential analysis tool for this complex type of data, and (3) the validation of the potential of nonstandardized protocols (in this case, the handgrip test) as alternatives to those currently carried out in clinical practice.

This work also presents some limitations. The fact that only two clinical features were considered to evaluate response at each time point (heart rate and basal septal velocity curve) limits the response characterization to these features. The low number of subjects in the study, owing in part to the rare nature of the disease, also limits the scope of the characterization of response. For a better/more detailed characterization, more features should be accounted for (e.g. regional deformation/velocity curves, flow information, etc.), as well as more cases. Furthermore, for a better assessment of the value of the proposed methodology and nonstandardized protocols, different populations, with different diseases in varying degrees should be considered. Nonetheless, despite these limitations, we believe that this work serves the purpose of illustrating the potential of the proposed methodology and nonstandardized stress testing.

5.2. Application II: Labour monitoring and decision making.

In a context of lacking evidence of positive impact of the current reference labour monitoring and decision support tool – the partograph – on outcome, and scepticism regarding the accuracy and generalizability of its central elements for diagnosis of (ab)normality in labour progression, the World Health Organization (WHO) has identified the

need to develop new evidence-based, personalized *Simplified, Effective, Labour Monitoring-to-Action* tools. In this application, we showed how unsupervised multiview dimensionality reduction (in particular, unsupervised MKL) could serve as the backbone of such a system. For the monitoring of a new patient, at each follow-up, the group of most similar "training" subjects (peers) is retrieved, and knowledge on their labour progress, interventions and outcomes is used to update the concept of "normal" labour progression and estimates of risk of intervention (and most likely timings). All these operations build upon a low-dimensional yet interpretable representation of the original data.

The main contribution of this work is the formulation of a new labour monitoring and decision support framework that overcomes the main limitations of the partograph and outperforms it in the prediction of meaningful labour events, while preserving interpretability, and its integration in a functional, scalable, web- and cloud-based prototype of a user-friendly clinical software tool (BCN-SELMA). The proposed approach also showed potential in the identification and understanding of practice differences/biases.

There is, however, room for improvement. On the one hand, the performance levels are still not ideal. Many steps, from data imputation to the estimation of the "normal" trajectory and risk estimates, were based on fairly simple methods that are likely not optimal. Future work should thus include the optimization of each step of the framework by exploring more sophisticated methodologies. On the other hand, validation is extremely challenging in terms of the prediction of actual risk of adverse outcome/"necessary" interventions, as with the currently available data we only have knowledge on what interventions were performed and the resulting outcomes, but no guarantee of causality. A better assessment of the value of the framework implies a wider prognostic evaluation. Future work should include the evaluation of our tool based on multiple studies/databases, as well as by clinicians. Regarding the implementation of the software prototype, the current priorities would be improving interpretability

(iterating over clinicians' feedback) and online performance.

Despite being based on a minimal implementation, and the validation limitations posed by the nature of the data, we believe that this work illustrates the potential of the proposed approach as basis for a monitoring and decision support tool.

5.3. Overall

We have developed tools based on unsupervised multiview dimensionality reduction (in particular, unsupervised MKL) for the analysis of "real-world" clinical longitudinal data, and illustrated how they could be used to: obtain simplified, interpretable representations of the data; discriminate between normal and abnormal trajectories and understand underlying pathophysiological mechanisms; and provide a basis for a personalized monitoring and decision support system.

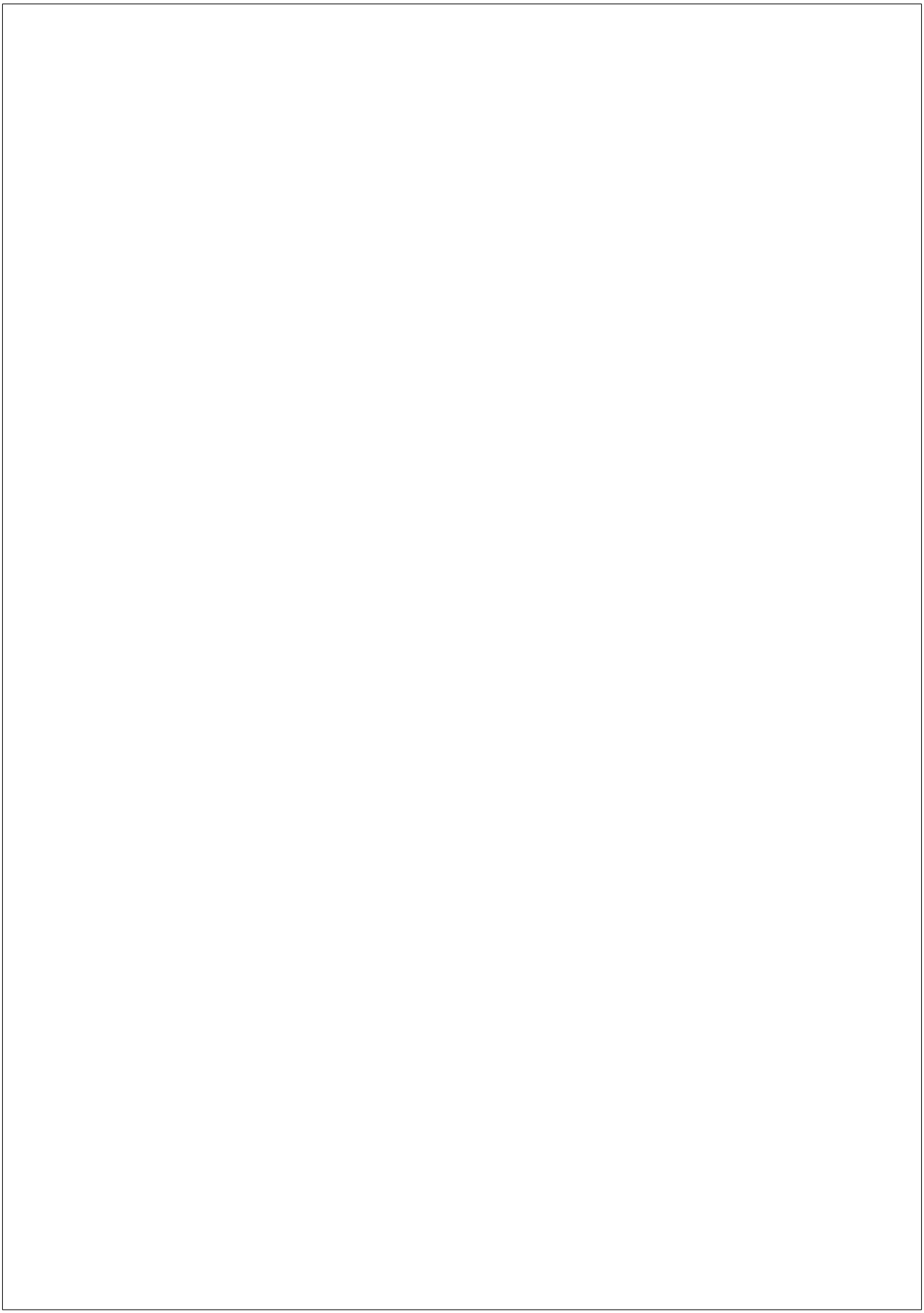
This was demonstrated while addressing two specific clinical problems – analysis of nonstandardized stress echocardiography and monitoring and decision support during labour. In both cases, this work has made valuable contributions, as described in 5.1 and 5.2. Nonetheless, for a better assessment of the value of the developed methodologies, further evaluations should be conducted with other studies/databases, and by clinicians.

We also presented a preliminary effort envisaging effective integration of the developed tools in a clinical environment – the web- and cloud-based prototype of BCN-SELMA. A necessary next step would be extensive evaluation and testing by clinicians, and feedback-based optimization of usability and interpretability. Technically, the priorities would be maturing the methodology (e.g. exploring more sophisticated implementations) towards better predictive performances and ensuring proper scalability and online (computational) performance.

Although applied to specific clinical problems, the developed tools – including the choice of MKL as the central algorithm – were designed

in a generic enough way to easily adapt to any other clinical problem (e.g. natural progression, therapy assessment, etc.) involving any type of data, provided there is an adequate way to express similarity through kernel functions, thereby expanding the scope of this thesis' contributions. With regard to MKL, optimizing the developed tools would also imply optimizing the choice of kernel types for the different types of data, a topic that was not exhaustively explored in this thesis. On the other hand, computational complexity of MKL is still a relevant limitation, and future research should include the development of methodologies to speed up computations. Future research should also include exploring different dimensionality reduction algorithms.

Despite the limitations listed across this chapter, we believe that the current work succeeds to showcase the potential of this type of approach in the analysis of "real-world" longitudinal clinical data.



Bibliography

[Ambrosino et al., 1995] Ambrosino, R., Buchanan, B., Cooper, G., and Fine, M. (1995). The use of misclassification costs to learn rule-based decision support models for cost-effective hospital admission strategies. *Proceedings of the Annual Symposium on Computer Application in Medical Care. Symposium on Computer Applications in Medical Care*, pages 304–8.

[Bach et al., 2004] Bach, F. R., Lanckriet, G. R. G., and Jordan, M. I. (2004). Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 6–. ACM.

[Bailit et al., 2005] Bailit, J., Dierker, L., Blanchard, M., and Mercer, B. (2005). Outcomes of women presenting in active versus latent phase of spontaneous labor. *Obstetrics and gynecology*, 105:77–9.

[Barcelona Supercomputing Center,] Barcelona Supercomputing Center. Nord III user's guide: System overview. `https://www.bsc.es/user-support/nord3.php#systemoverview`. Accessed: 2020-03-02.

[Bemdt and Clifford, 1994] Bemdt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series.

[Bermanis et al., 2013] Bermanis, A. et al. (2013). Multiscale data sampling and function extension. *Applied and Computational Harmonic Analysis*, 34(1):15–29.

[Betrán et al., 2015] Betrán, A., Torloni, M., J Zhang, J., Gülmezoglu, A., and Zongo, A. (2015). Who statement on caesarean section rates. *BJOG: An International Journal of Obstetrics Gynaecology*, 123.

[Betrán et al., 2018] Betrán, A. P., Temmerman, M., Kingdon, C., Mohiddin, A., Opiyo, N., Torloni, M. R., Zhang, J., Musana, O., Wanyonyi, S. Z., Gülmezoglu, A. M., and Downe, S. (2018). Interventions to reduce unnecessary caesarean sections in healthy women and babies. *The Lancet*, 392(10155):1358 – 1368.

[Boatin et al., 2018] Boatin, A. A., Schlotheuber, A., Betran, A. P., Moller, A.-B., Barros, A. J. D., Boerma, T., Torloni, M. R., Victora, C. G., and Hosseinpoor, A. R. (2018). Within country inequalities in caesarean section rates: observational study of 72 low and middle income countries. *BMJ*, 360.

[Bonet et al., 2019] Bonet, M., Oladapo, O., Souza, J. P., and Gülmezoglu, A. M. (2019). Diagnostic accuracy of the partograph alert and action lines to predict adverse birth outcomes: a systematic review. *BJOG: An International Journal of Obstetrics & Gynaecology*.

[Borchers, 1999] Borchers, B. (1999). Csdp, a c library for semidefinite programming. *Optimization Methods and Software*, 11(1-4):613–623.

[Burke et al., 2017] Burke, N., Burke, G., Breathnach, F., Mcauliffe, F., Morrison, J., Turner, M., Dornan, S., Higgins, J., Cotter, A., Geary, M., Mcparland, P., Daly, S., Cody, F., Dicker, P., Tully, E., and Malone, F. (2017). Prediction of cesarean delivery in the term nulliparous woman: Results from the prospective multi-center genesis study. *American Journal of Obstetrics and Gynecology*, 216.

[Campillo-Artero et al., 2018] Campillo-Artero, C., Serra-Burriel, M., and Calvo-Pérez, A. (2018). Predictive modeling of emergency cesarean delivery. *PloS one*, 13(1):e0191248.

[Caruana et al., 2015] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 1721–1730, New York, NY, USA. Association for Computing Machinery.

[Chen et al., 2004] Chen, G., Uryasev, S., and Young, T. K. (2004). On prediction of the cesarean delivery risk in a large private practice. *American Journal of Obstetrics and Gynecology*, 191(2):616 – 623.

[Chuma et al., 2014] Chuma, C., Kihunrwa, A., Matovelo, D., and Mahendeka, M. (2014). Labor management and obstetric outcomes among pregnant women admitted in latent phase compared to active phase of labor at bugando medical centre in tanzania. *BMC pregnancy and childbirth*, 14:68.

[Cikes et al., 2019] Cikes, M., Sanchez-Martinez, S., Claggett, B., Duchateau, N., Piella, G., Butakoff, C., Pouleur, A. C., Knappe, D., Biering-Sørensen, T., Kutyifa, V., Moss, A., Stein, K., Solomon, S. D., and Bijnens, B. (2019). Machine learning-based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy. *European Journal of Heart Failure*, 21(1):74–85.

[Daemen and De Moor, 2009] Daemen, A. and De Moor, B. (2009). Development of a kernel function for clinical data. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5913–5917.

[Dagum and Menon, 1998] Dagum, L. and Menon, R. (1998). Openmp: an industry standard api for shared-memory programming. *IEEE Computational Science and Engineering*, 5(1):46–55.

[Davidavicius et al., 2003] Davidavicius, G. et al. (2003). Can regional strain and strain rate measurement be performed during both dobutamine and exercise echocardiography, and do regional

deformation responses differ with different forms of stress testing? *Journal of the American Society of Echocardiography*, 16(4):299–308.

[Duchateau et al., 2013] Duchateau, N. et al. (2013). Adaptation of multiscale function extension to inexact matching: Application to the mapping of individuals to a learnt manifold. *Lecture Notes in Computer Science*, pages 578–586.

[Epanechnikov, 1969] Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1):153–158.

[Friedman, 1954] Friedman, E. A. (1954). The graphic analysis of labor. *American Journal of Obstetrics and Gynecology*, 68(6):1568 – 1575.

[Gropp et al., 1996] Gropp, W., Lusk, E., Doss, N., Skjellum, A., Mathematics, Science, C., and Univ., M. S. (1996). A high-performance, portable implementation of the mpi message passing interface standard. *Parallel Comput.*, 22.

[Guennebaud et al., 2010] Guennebaud, G., Jacob, B., et al. (2010). Eigen v3. http://eigen.tuxfamily.org.

[Harper et al., 2013] Harper, L., Odibo, A., Macones, G., and Cahill, A. (2013). Predicting cesarean in the second stage of labor. *American Journal of Perinatology*, 30.

[Hearst et al., 1998] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28.

[Helfant et al., 1971] Helfant, R. H. et al. (1971). Effect of sustained isometric handgrip exercise on left ventricular performance. *Circulation*, 44(6):982–993.

[Holmes et al., 2001] Holmes, P., Oppenheimer, L. W., and Wen], S. W. (2001). The relationship between cervical dilatation at initial presentation in labour and subsequent intervention. *British Journal of Obstetrics and Gynaecology*, 108(11):1120 – 1124.

[Hotelling, 1936] Hotelling, H. (1936). Relations between two sets of variables. *Biometrika*, 28(3-4):321–377.

[Janssen et al., 2017] Janssen, P. A., Stienen, J. J. C., Brant, R. F., and Hanley, G. E. (2017). A predictive model for cesarean among low-risk nulliparous women in spontaneous labor at hospital admission. *Birth*, 44:21–28.

[Kivowitz et al., 1971] Kivowitz, C., Parmley, W. W., Donoso, R., Marcus, H., Ganz, W., and Swan, H. J. C. (1971). Effects of isometric exercise on cardiac performance. *Circulation*, 44(6):994–1002.

[Levine et al., 2018] Levine, L. D., Downes, K. L., Parry, S., Elovitz, M. A., Sammel, M. D., and Srinivas, S. K. (2018). A validated calculator to estimate risk of cesarean after an induction of labor with an unfavorable cervix. *American Journal of Obstetrics and Gynecology*, 218(2):254.e1 – 254.e7.

[Li et al., 2018] Li, Y., Yang, M., and Zhang, Z. M. (2018). A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1.

[Lin YY, 2011] Lin YY, Liu TL, F. C. (2011). Multiple kernel learning for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1147–60.

[Mikolajczyk et al., 2016] Mikolajczyk, R., Jun, Z., Grewal, J., Chan, L., Petersen, A., and Gross, M. (2016). Early versus late admission to labor affects labor progression and risk of cesarean section in nulliparous women. *Frontiers in Medicine*, 3.

[National Institute for Health and Clinical Excellence., 2007] National Institute for Health and Clinical Excellence. (2007). Intrapartum care: care of healthy women and their babies during childbirth. *NICE clinical guideline 55, 2007. Available at guidance.nice.org.uk/cg55.*

[Neal et al., 2014] Neal, J., Lamp, J., Buck, J., Lowe, N., Gillespie, S., and Ryan, S. (2014). Outcomes of nulliparous women with spontaneous labor onset admitted to hospitals in preactive versus active labor. *Journal of midwifery women's health*, 59.

[Nogueira et al., 2020a] Nogueira, M., Craene, M. D., Sanchez-Martinez, S., Chowdhury, D., Bijnens, B., and Piella, G. (2020a). Analysis of nonstandardized stress echocardiography sequences using multiview dimensionality reduction. *Medical Image Analysis*, 60:101594.

[Nogueira et al., 2020b] Nogueira, M., Piella, G., Craene, M. D., Yagüe, C., Sanchez-Martinez, S., Martí, P., Bonet, M., Oladapo, O. T., and Bijnens, B. (2020b). A personalised approach for effective labour monitoring based on machine learning of women's similarity and optimal temporal progression. *To be submitted.*

[Nogueira et al., 2017] Nogueira, M., Piella, G., Sánchez Martínez, S., Langet, H., Saloux, E., Bijnens, B., and De Craene, M. (2017). Characterizing patterns of response during mild stress-testing in continuous echocardiography recordings using a multiview dimensionality reduction technique. pages 502–513.

[Oladapo et al., 2015] Oladapo, O., Souza, J. P. D. d., Bohren, M. A., Tunçalp, , Vogel, J. P., Fawole, B., Mugerwa, K., and Gulmezoglu, A. M. (2015). WHO better outcomes in labour difficulty (BOLD) project: innovating to improve quality of care around the time of childbirth. *Reproductive Health.*

[Oladapo et al., 2018] Oladapo, O. T., Souza, J. P., Fawole, B., Mugerwa, K., Perdoná, G., Alves, D., Souza, H., Reis, R., Oliveira-Ciabati, L., Maiorano, A., Akintan, A., Alu, F. E., Oyeneyin, L., Adebayo, A., Byamugisha, J., Nakalembe, M., Idris, H. A., Okike, O., Althabe, F., Hundley, V., Donnay, F., Pattinson, R., Sanghvi, H. C., Jardine, J. E., Özge Tunçalp, Vogel, J. P., Stanton, M. E., Bohren, M., Zhang, J., Lavender, T., Liljestrand, J., ten Hoope-Bender, P., Mathai, M., Bahl, R., and Gülmezoglu, A. M. (2018). Progression of the first stage of spontaneous labour: A prospective cohort study in two sub-saharan african countries. *PLOS Medicine*, 15(1):e1002492.

[Peressutti et al., 2017] Peressutti, D., Sinclair, M., Bai, W., Jackson, T., Ruijsink, J., Nordsletten, D., Asner, L., Hadjicharalambous, M., Rinaldi, C. A., Rueckert, D., and King, A. P. (2017). A framework for combining a motion atlas with non-motion information to learn clinically useful biomarkers: Application to cardiac resynchronisation therapy response prediction. *Medical Image Analysis*, 35:669 – 684.

[Puyol-Antón et al., 2019] Puyol-Antón, E., Ruijsink, B., Gerber, B., Amzulescu, M. S., Langet, H., De Craene, M., Schnabel, J. A., Piro, P., and King, A. P. (2019). Regional multi-view learning for cardiac motion analysis: Application to identification of dilated cardiomyopathy patients. *IEEE Transactions on Biomedical Engineering*, 66(4):956–966.

[Robson et al., 2015] Robson, M., Murphy, M., and Byrne, F. (2015). Quality assurance: The 10-group classification system (robson classification), induction of labor, and cesarean delivery. *International Journal of Gynecology Obstetrics*, 131:S23 – S27. World Report on Women's Health 2015: The unfinished agenda of women's reproductive health.

[Rudin et al., 1992] Rudin, L. I., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259 – 268.

[Sanchez-Martinez et al., 2018] Sanchez-Martinez, S., Duchateau, N., Erdei, T., Kunszt, G., Aakhus, S., Degiovanni, A., Marino, P., Carluccio, E., Piella, G., Fraser, A. G., and Bijnens, B. H. (2018). Machine learning analysis of left ventricular function to characterize heart failure with preserved ejection fraction. *Circulation: Cardiovascular Imaging*, 11(4).

[Sanchez-Martinez et al., 2017] Sanchez-Martinez, S. et al. (2017). Characterization of myocardial motion patterns by unsupervised multiple kernel learning. *Medical Image Analysis*, 35:70–82.

[Sanchez-Martinez et al.,] Sanchez-Martinez, S., Piella, N. D. G., and Butakoff, C. Unsupervised multiple kernel learning. `https://github.com/bcnmedtech/unsupervised_multiple_kernel_learning`.

[Souza et al., 2019] Souza, H., Perdoná, G., Marcolin, A., Oyeneyin, L., Oladapo, O., Mugerwa, K., and Souza, J. (2019). Development of caesarean section prediction models: secondary analysis of a prospective cohort study in two sub-saharan african countries. *Reproductive Health*, 16.

[Souza et al., 2015] Souza, J., Oladapo, O., Bohren, M., Mugerwa, K., Fawole, B., Moscovici, L., Perdona, G., Oliveira-Ciabati, L., Vogel, J., Tunalp, Ö., Zhang, J., Bahl, R., Gülmezoglu, A., Agrawal, P., Alves, D., Armbruster, D., Donnay, F., Hofmeyr, J., Hoope-Bender, P., Hundley, V., Kyaddondo, D., Lavender, T., Lewin, S., Miettinen, S., Olutayo, L., Pattinson, B., Ramsey, K., Stanton, M., and Sanghvi, H. (2015). The development of a simplified, effective, labour monitoring-to-action (SELMA) tool for better outcomes in labour difficulty (BOLD): study protocol. *Reproductive Health*, 12.

[Souza et al., 2018] Souza, J., Oladapo, O., Fawole, B., Mugerwa, K., Reis, R., Barbosa-Junior, F., Oliveira-Ciabati, L., Alves, D., and Gülmezoglu, A. (2018). Cervical dilatation over time is a poor predictor of severe adverse birth outcomes: a diagnostic accuracy study. *BJOG: An International Journal of Obstetrics & Gynaecology*, 125(8):991–1000.

[Strauss et al., 2013] Strauss, K. A., DuBiner, L., Simon, M., Zaragoza, M., Sengupta, P. P., Li, P., Narula, N., Dreike, S., Platt, J., Procaccio, V., Ortiz-González, X. R., Puffenberger, E. G., Kelley, R. I., Morton, D. H., Narula, J., and Wallace, D. C. (2013). Severity of cardiomyopathy associated with adenine nucleotide translocator-1 deficiency correlates with mtdna haplogroup. *Proceedings of the National Academy of Sciences*, 110(9):3453–3458.

[Velasco et al., 1997] Velasco, M. et al. (1997). The cold pressor test. *American Journal of Therapeutics*, 4(1):34–38.

[Virtanen et al., 2020] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E. W., Vand erPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and Contributors, S. . . (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

[Voigt, 2003] Voigt, J.-U. (2003). Strain-rate imaging during dobutamine stress echocardiography provides objective evidence of inducible ischemia. *Circulation*, 107(16):2120–2126.

[Ward Jr., 1963] Ward Jr., J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

[Wold, 1985] Wold, H. (1985). *Partial Least Squares*. John Wiley & Sons, Inc.

[World Health Organization,] World Health Organization. Global health observatory data repository: Births by caesarean section – data by country. `http://apps.who.int/gho/data/node.main.BIRTHSBYCAESAREAN?lang=en`. Accessed: 2019-08-21.

[World Health Organization., 2014] World Health Organization. (2014). WHO recommendations on augmentation of labour.

[World Health Organization, 2015] World Health Organization (2015). Trends in maternal mortality: 1990-2015: estimates from WHO, UNICEF, UNFPA, world bank group and the united nations population division: executive summary. Technical documents.

[Xu et al., 2013] Xu, C., Tao, D., and Xu, C. (2013). A survey on multi-view learning. *CoRR*, abs/1304.5634.

[Yang et al., 2017] Yang, F., Bohren, M., Kyaddondo, D., Musibau Ayoade, T., Olutayo, A., Oladapo, O., Souza, J., Gülmezoglu, A., Mugerwa, K., and Fawole, B. (2017). Healthcare providers' perspectives on labor monitoring in nigeria and uganda: A qualitative study on challenges and opportunities. *International Journal of Gynecology & Obstetrics*, 139 Suppl 1.

Publications

Journals

- **M. Nogueira**, M. De Craene, S. Sanchez-Martinez, D. Chowdhury, B. Bijnens, G. Piella. Analysis of nonstandardized stress echocardiography sequences using multiview dimensionality reduction. *Medical Image Analysis*, 60:101594.
- **M. Nogueira**, G. Piella, M. De Craene, C. Yagüe, S. Sanchez-Martinez, P. Martí, M. Bonet, O.T. Oladapo, B. Bijnens. A personalised approach for effective labour monitoring based on machine learning assessing women's similarity and optimal temporal progression. *Submitted to Nature Methods.*
- **M. Nogueira**, C. Yagüe, G. Piella, M. De Craene, S. Sanchez-Martinez, P. Martí, M. Bonet, O.T. Oladapo, B. Bijnens. BCN-SELMA: A Simplified, Effective, Labour Monitoring-to-Action tool, based on Interpretable Machine Learning. *In preparation.*

Conferences

- **M. Nogueira**, G. Piella, S. Sanchez-Martinez, H. Langet, E. Saloux, B. Bijnens, M. De Craene. Characterizing patterns of response during mild stress-testing in continuous echocardiography recordings using a multiview dimensionality reduction technique.

Functional Imaging and Modelling of the Heart. Lecture Notes on Computer Science, volume 10263, 2017.

Funding

This work was supported by:

- The European Union's Horizon 2020 Programme for Research and Innovation, under grant agreement No. 642676 (CardioFunXion);
- The Spanish Ministry of Economy and Competitiveness (grant TIN2014-52923-R; Maria de Maeztu Units of Excellence Programme - MDM-2015-0502);
- The Fundació La Marató de TV3 (No. 20154031);
- The Bill Melinda Gates Foundation (Grant OPP1084318: https://www.gatesfoundation.org/How-We-Work/Quick-Links/Grants-Databaseq/k=OPP1084318);
- The United States Agency for International Development (USAID);
- The UNDP-UNFPA-UNICEF-WHO-World Bank Special Programme of Research, Development and Research Training in Human Reproduction (HRP), a cosponsored program executed by the World Health Organization (WHO).